

SoK: Reconstruction Attacks on Synthetic Tabular Data (Insights from Winning the NIST CRC)

Steven Golob¹, Sikha Pentyla¹, Martine De Cock^{1,2}

¹School of Engineering and Technology, University of Washington Tacoma, Tacoma, WA, USA

²Department of Mathematics, Computer Science, and Statistics, Ghent University, Gent, Belgium
golobs@uw.edu

Abstract

Synthetic data is increasingly promoted as a privacy-preserving substitute for releasing sensitive tabular records, yet its central adversarial threat (*reconstruction*, the recovery of an individual’s hidden attribute values from a synthetic release and a handful of known quasi-identifiers) has been studied only in scattered, hard-to-compare settings. We present the first systematization of reconstruction (equivalently, attribute inference) attacks on de-identified and synthetic tabular data. We contribute a taxonomy that organizes attacks by the structure they exploit; the most systematic empirical evaluation to date, pitting fourteen attacks against nine synthetic data generation (SDG) methods across five benchmark datasets; and a set of new attacks that fill gaps in the taxonomy, one of which (CoBP-RA) is the strongest attack we measure. Crucially, we introduce a methodology for interpreting what attack success means: a memorization test that distinguishes reconstruction of the population distribution from memorization of training records, and a reduction that places reconstruction and membership inference on a single comparable scale. Our findings: the choice of SDG method governs risk far more than the choice of attack; differential privacy protects mainly at small budgets ($\epsilon \lesssim 1$), above which protection plateaus, bounded by the synthesizer’s capacity rather than its noise; de-identification methods are the most exposed; and most reconstruction reflects distributional structure rather than memorization, concentrating individual risk on atypical records. The attacks and infrastructure are externally validated by our first-place finish among all red teams in the 2025 *National Institute of Standards and Technology* (NIST) Collaborative Research Cycle.

Keywords

privacy, synthetic data, de-identification, reconstruction attack, attribute inference

1 Introduction

Synthetic data promises to resolve a central tension in data sharing: produce a dataset that reproduces the statistical structure of sensitive records, and publish it for downstream users to analyze without exposing any real individual [34]. The promise has drawn

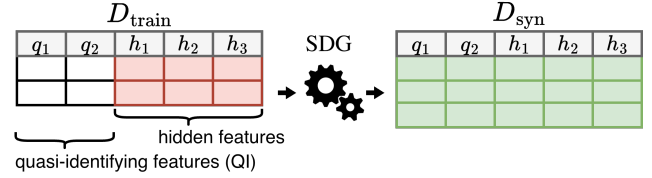


Figure 1: The reconstruction threat model (synthetic-data branch of Figure 2; the de-identified branch is identical, with SDG replaced by de-identification). The SDG process turns the private training table into the synthetic release (green rows). Every training record is a target: the adversary sees the records’ quasi-identifiers (white columns) plus the synthetic data, and fills in their hidden attributes (red columns).

substantial institutional investment [36]. National statistical agencies [8], hospital systems [67], and the financial sector [5] are evaluating synthetic data release as a successor to traditional disclosure-limitation techniques such as k -anonymization [62] and record suppression [55], and as a path to compliance under frameworks like HIPAA Expert Determination [64].

Whether that promise of privacy survives adversarial scrutiny is an empirical question, and the sharpest test of it is the **reconstruction attack**. Given a released synthetic dataset and a partial view of a target individual—a set of *quasi-identifying* features, or **quasi-identifiers** (QI): attributes such as age, sex, or ZIP code that an adversary can plausibly obtain from public records, voter rolls, or social-media profiles [49, 61]—the adversary attempts to recover that individual’s remaining, hidden attributes (Figure 1: the adversary knows the *white* quasi-identifier columns and must fill in the *red* hidden columns). Reconstruction thus completes a partial record, predicting sensitive values such as income, disease status, or household wealth.¹ It is strictly more consequential than *membership inference* [57] (the de facto privacy audit for synthetic data [3, 28, 58]), which asks only *whether* a record was used in training, never *what* that record contains.

Reconstruction attacks divide by the *artifact* the adversary holds (Figure 2): a trained model (*model inversion* [24]), released aggregate statistics (*reconstruction from tabulations* [18]), or a published de-identified or synthetic table. We study the last: the setting of the NIST Collaborative Research Cycle,² where both pose one threat model and attack surface (Section 1.3).

¹“Attribute inference” is sometimes reserved in prior work for predicting a single withheld attribute [72]. We use “reconstruction” in the broader sense of predicting all of a target record’s hidden attributes at once. Under the threat model we study (Section 1.3), attribute inference and reconstruction are the same problem, and we use the terms interchangeably.

²https://pages.nist.gov/privacy_collaborative_research_cycle/

This work is licensed under the Creative Commons Attribution 4.0 International License. To view a copy of this license visit <https://creativecommons.org/licenses/by/4.0/> or send a letter to Creative Commons, PO Box 1866, Mountain View, CA 94042, USA.



Under Review, 1–27

© YYYY Copyright held by the owner/author(s).

<https://doi.org/XXXXXXX.XXXXXXX>

Why systematize? Despite its centrality to disclosure risk, reconstruction from synthetic tabular data has, to our knowledge, never been systematized. The literature is a patchwork: studies fix a single attack method [4, 58] or a single privacy metric [27, 33], and report numbers that cannot be placed side by side. Basic questions therefore remain open. Which attacks actually work, and against which SDG methods? Does the adversary’s sophistication or the data holder’s choice of SDG method drive risk? When an attack succeeds, has it memorized a training record or merely learned the population distribution? And how can a reconstruction attack be compared against a membership inference attack, when the two are scored on different scales? No single prior study answers these; doing so requires a unifying taxonomy of reconstruction attacks, a broad and controlled evaluation, and a methodology for interpreting what a high attack score *means*. This paper supplies all three through the following *main contributions*:

- **A taxonomy of reconstruction attacks** (Section 2, Figure 2), to our knowledge the first systematization of this area. It organizes attacks by the structure they exploit, separating those that predict each hidden feature *in isolation* [4, 58] from those that exploit *correlations among hidden features* (autoregressive, message-passing, and joint-generative).
- **The most systematic empirical evaluation to date** (Section 3): fourteen attacks against nine SDG methods (the seven from the NIST CRC plus a GAN- and a diffusion-based generator) spanning differentially private (DP) synthesizers, non-private deep generative models, and de-identification, evaluated across five datasets with quasi-identifier sets varied in composition and size (3 to 16 features). This substantially extends the NIST CRC, which fixed a single dataset and QI configuration.
- **New attacks that populate sparse branches** of the taxonomy, offered as an exploration of the design space rather than an exhaustive enumeration: **CoBP-RA**, which runs belief propagation of per-feature random-forest posteriors over a learned graphical model and is the single strongest attack in our study; and a family that recasts reconstruction as QI-conditioned synthesis: **CondMST** [45] and the diffusion-based **CondDDPM** [38] and **CondRePaint**, the latter porting image-inpainting *RePaint* [40] to tables. CoBP-RA improves on the strongest prior attack across every categorical dataset we test; the graphical-model and diffusion attacks are competitive but do not consistently beat well-tuned per-feature classifiers (Section 2). Multiple branches remain open for future attacks.
- **A methodology for interpreting attack success** (Section 4). Measured only against training targets (as in the literature and the NIST CRC), a high score is ambiguous: it may reflect memorization of training records (a breach) or merely a well-modeled population distribution. Our memorization test runs each attack against training and held-out targets from the same population; the gap (an attribute analogue of the train-versus-holdout gap behind membership advantage [72], here per hidden feature) separates the two. We furthermore propose a reduction that casts a reconstruction attack as a membership inference attack (*RA-as-MIA*), placing both paradigms on a single axis.
- **Policy-relevant guidance** for data holders and auditors, synthesizing the above into recommendations (Sections 3 and 5).

External validation. The attacks and infrastructure are not benchmarked only in-house: the same methodology placed first among all red teams in the 2025 NIST Privacy Collaborative Research Cycle (CRC)³, a structured red-team/blue-team program whose privacy-measurement results inform policy at federal statistical agencies; the US Census Bureau, for instance, adopted differential privacy for its 2020 census data releases [1].

Main findings:

- **The generator, not the adversary, is the dominant lever on risk.** Reconstruction accuracy varies far more with the SDG method than with the attack; a more protective generator lowered risk more than any attack refinement we tried (Section 2).
- **Differential privacy works, but chiefly at small budgets.** Protection accrues at $\epsilon \lesssim 1$; above it the marginal DP synthesizers are limited not by noise but by the pairwise structure they can represent, so more budget buys little (Section 3).
- **De-identification is the most exposed family.** Cell suppression and rank swapping leave more reconstructable structure than even high-fidelity generative models (Section 3).
- **Adversary knowledge quality beats quantity.** The *composition* of the quasi-identifying features shapes risk more than its size (Section 3).
- **Most reconstruction is distributional, not memorization.** Our memorization test (Section 4) separates the two: attacks succeed largely by reproducing population structure the release preserves rather than memorizing individuals. It reframes what a high score means and reveals where risk concentrates: on atypical, demographically-rare records, who can face several times the average risk under high-fidelity synthesis.
- **Reconstruction and membership inference agree in aggregate but diverge per record.** The two rank SDG methods almost identically, yet succeed on substantially different individuals (about a third of records give discordant signals), so neither subsumes the other and a thorough audit needs both (Section 4.2).

1.1 Formal Problem Description

Let $D_{\text{train}} = \{x_1, \dots, x_n\}$ be a tabular dataset over n individuals, each record a vector of p features, each feature either categorical or continuous. We write $\mathcal{F}$ for the set of features, $|\mathcal{F}| = p$. A **synthetic data generator** (SDG) $\mathcal{G}$ takes D_{train} as input and produces a release $D_{\text{syn}} = \mathcal{G}(D_{\text{train}})$ (we write D_{syn} for both synthetic and de-identified releases, since we treat them identically) intended to preserve the statistical structure of the original without containing real records.

We partition features into two disjoint sets: the **quasi-identifier** (QI) $Q \subset \mathcal{F}$ comprising features assumed observable by the adversary from auxiliary sources [49, 61] (the *white* columns of Figure 1), and the **hidden features** $H = \mathcal{F} \setminus Q$ that the adversary wishes to infer (the *red* columns). A **reconstruction attack** is a function

$$f : (D_{\text{syn}}, x_i^Q) \mapsto \hat{x}_i^H,$$

mapping the synthetic release and a target’s observed quasi-identifiers to a prediction of the target’s hidden attributes.

³https://pages.nist.gov/privacy_collaborative_research_cycle/pages/red_team.html

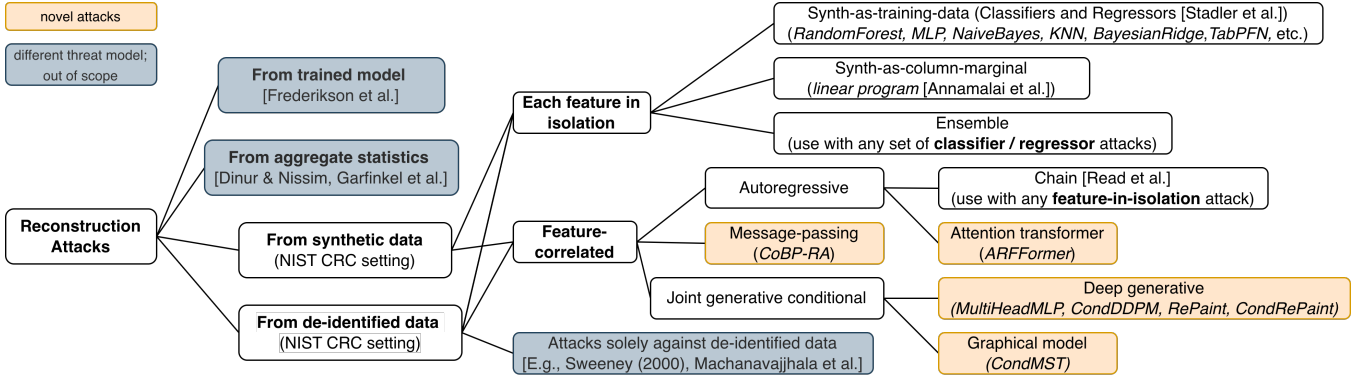


Figure 2: Taxonomy of reconstruction attacks. We study the two lower branches (reconstruction from de-identified microdata and from synthetic data), which share the NIST CRC quasi-identifier threat model; we treat them identically and merge their (here identical) attack subtrees (a modeling choice: the de-identified branch additionally admits attacks on *retained real records*, Section 2.1). Within the merged subtree, per-feature *in-isolation* attacks fit one model per hidden feature (a classifier or regressor), while *feature-correlated* attacks exploit dependencies among the hidden features, the category our new attacks target (Section 2.2). The top two branches (model inversion and reconstruction from tabulations) assume a different release artifact and are out of scope.

Scoring: rarity-weighted reconstruction advantage (R_{adv}). Naïve accuracy rewards predicting the mode, a trivially achievable but uninformative baseline. We use a *rarity-weighted* score: a correct prediction of rare value v earns weight $n/\text{count}(v, D_{\text{train}})$ (where $\text{count}(v, D_{\text{train}})$ is the number of training records in D_{train} taking value v on feature h), reflecting the greater harm of inferring information the adversary could not have guessed. The score for inferring hidden feature h is:

$$R_{adv}(f, h) = \frac{\sum_i \mathbf{1}[\hat{x}_i^h = x_i^h] \cdot n/\text{count}(x_i^h, D_{\text{train}})}{\sum_i n/\text{count}(x_i^h, D_{\text{train}})} \times 100. \quad (1)$$

Under this measure a mode-baseline attack scores exactly $100/C\%$ regardless of the distribution’s shape, where C is the number of distinct values; a perfect attack scores 100%. This was also the official scoring metric for the NIST Privacy CRC competition. We report R_{adv} as a percentage; overall attack scores are the mean R_{adv} across all hidden features.

1.2 Differential Privacy

Differential privacy (DP) [19, 20] is the principal formal privacy guarantee studied here. A randomized mechanism $\mathcal{M}$ is ϵ -**differentially private** if, for neighboring datasets D, D' (i.e., datasets that differ in one record) and all subsets S of the range of $\mathcal{M}$: $\Pr[\mathcal{M}(D) \in S] \leq e^\epsilon \Pr[\mathcal{M}(D') \in S]$. Smaller ϵ provides stronger privacy; an ϵ -DP SDG guarantees that no individual’s presence in D_{train} shifts the distribution of D_{syn} by more than e^ϵ . The bound is worst-case and does not directly specify empirical reconstruction accuracy; we measure R_{adv} across $\epsilon \in \{10^{-1}, \dots, 10^3\}$ for the DP SDG methods MST [43] and AIM [44] in Section 3.

1.3 Threat Model

Setting and adversary. A data holder publishes the release D_{syn} in place of D_{train} ; the adversary targets a specific training individual x_i and aims to recover its hidden attributes x_i^H . The adversary observes

only D_{syn} and the target’s quasi-identifiers x_i^Q : never the original data, the generator’s parameters, seed, or weights (the attack is fully black-box), nor any other individual’s record. This is not a single, static threat model: the quasi-identifier set Q is the knob that sets how much the adversary knows, and we vary it in both size and composition (Section 3.3), so that a larger or more informative Q instantiates a correspondingly stronger adversary.

What we do and do not study. We study the lower two branches of Figure 2: reconstruction from de-identified microdata and from synthetic data, treated identically. Out of scope are model inversion and reconstruction-from-tabulations (which assume a different release artifact: a trained model or aggregate statistics), along with attacks that jointly exploit multiple target records, model extraction, and white-box attacks on SDG internals. Ours is the realistic black-box, single-record auditing setting relevant to published releases and structured compliance evaluations.

Paper organization. Section 2 presents the attack taxonomy and our main cross-method benchmark. Section 3 examines how the SDG method and its privacy budget, the data domain (demographic, health, financial, and continuous valued datasets), the dataset size, and quasi-identifier design shape reconstruction risk. Section 4 interrogates what attack success *means*: separating memorization from distributional inference, comparing reconstruction against membership inference, and measuring disparate impact across sub-populations. Section 5 translates these findings into auditing and policy guidance. Two enhancements that aided our competition submission but yield no systematic gain in controlled experiments (chaining and ensembling) are summarized briefly here and developed, with their oracle analyses, in the appendix.

2 A Taxonomy of Reconstruction Attacks

2.1 The Reconstruction-Attack Landscape

Reconstruction attacks form a family whose members we distinguish by the *artifact* the adversary has access to (Figure 2, top level).

We survey the family in full before narrowing to the two branches this paper studies.

Model inversion (out of scope). The first branch attacks a *trained model*: model inversion [24] exploits a classifier’s confidence scores to recover sensitive training attributes. Zhao et al. [75] argue that exact attribute inference is largely infeasible (only an approximate, imputation-like recovery succeeds), and Jayaraman and Evans [35] reduce many attribute-inference attacks to ordinary imputation; leverage comes from feature correlations, not a membership signal. A related federated-learning line, the *attribute reconstruction attack* (ARA), recovers training-data attributes from shared model gradients [14, 41]; despite the shared name it likewise assumes access to model internals rather than a published table. We borrow the imputation lens of these attacks, but our adversary never observes a model or its gradients, placing this entire branch out of scope.

Reconstruction from aggregate statistics (out of scope). A second, theoretically deep branch attacks *released tabulations*. Dinur and Nissim [18] proved that answering too many counting queries too accurately permits near-complete reconstruction, the result that motivated differential privacy. Garfinkel et al. [26] traced this threat to U.S. Census releases, informing the Bureau’s adoption of DP for 2020; Dick et al. [17] reconstructed and confidence-ranked individual census records; and Cohen et al. [15] characterized when reconstruction from aggregates is feasible, while Dwork et al. [21] survey the broader landscape and Muralidhar and Domingo-Ferrer [48] caution that reconstruction in practice is often harder than the theory implies. Our results are the generative-setting counterpart: reconstruction is consistently above chance, but its ceiling is set by the structure the release preserves (echoing Cohen et al.’s feasibility boundary), not by adversary effort.

Reconstruction from synthetic data (in scope). This branch attacks *synthetic data*: records drawn from a generative model fit to the original. Stadler et al. [58] framed attribute inference as a game in which a machine learning (ML) adversary trained on the synthetic release predicts withheld attributes of real records, concluding that synthetic data rarely provides strong anonymization. Houssiau et al. [33] (TAPAS) and Giomi et al. [27] (Anonymeter) assemble attack toolkits for auditing synthetic data (TAPAS spanning membership and attribute inference, Anonymeter the GDPR risks of singling-out, linkability, and attribute inference), but both are evaluation frameworks rather than systematic comparisons of which generators leak most. Ganev and De Cristofaro [25] show that the similarity-based privacy scores marketed by synthetic-data vendors badly understate real risk. Each of these works fixes a single attacker, threat model, or metric; none organizes the attack space or compares attacks on a common footing: the gap this paper closes.

Reconstruction from de-identified microdata (in scope). This branch attacks *de-identified record-level data*: a table whose identifiers have been removed or perturbed by suppression, generalization, or swapping. The re-identification literature established that such releases leak. Sweeney [61] showed that quasi-identifiers alone re-identify most individuals and proposed k -anonymity; Machanavajjhala et al. [42] then showed k -anonymity still permits *attribute* disclosure when a quasi-identifier group is homogeneous, motivating ℓ -diversity. We attack de-identified releases in this spirit,

but score attribute reconstruction rather than re-identification; the retained-record attacks unique to de-identification (re-identification from quasi-identifiers [61] and these homogeneity / background-knowledge attacks on k -anonymity) have no analog against fully synthetic data and are themselves out of scope (Figure 2).

Two branches, one attack surface. Because de-identified and synthetic releases present the adversary with the same information (a published record-level table plus the target’s QI), their attack methodologies coincide, and we merge the two subtrees of Figure 2.⁴ Both were the setting of the NIST CRC, where our submission placed first (Appendix E.7). We attack nine generators spanning three paradigms. Seven are those evaluated in the NIST CRC: the differentially private marginal synthesizers MST [43] and AIM [44] (cf. PrivBayes [74]); the deep generative models TVAE [70] and ARF [68]; and the de-identification methods Synthpop [52], rank swapping [47], and cell suppression [63]. We add two more to broaden coverage of the dominant deep-generative families: the GAN-based CTGAN [70] and the diffusion-based TabDDPM [38]. Full descriptions are deferred to Appendix A.2.

2.2 How Attacks on Synthetic Data Exploit Structure

Every attack we study maps a release D_{syn} and a target’s QI to a guess of the hidden attributes; they differ in *how much joint structure they exploit* (Figure 2, lower subtree). One family reconstructs each hidden feature *in isolation*; the other exploits the *correlations among hidden features*. We describe both, then benchmark them.

Reference points (not attacks). Three trivial procedures fix a scale for the rarity-weighted score R_{adv} , which (unlike the AUC of a binary test, where 0.5 is always chance) shifts with the dataset, the QI, and the generator. For each hidden feature h we (i) impute the **mode** (categorical) or **mean** (continuous) of $D_{\text{syn}}[h]$; (ii) draw a **random** value from $D_{\text{syn}}[h]$; or (iii) **copy** $D_{\text{syn}}[h]$ row-for-row (using $D_{\text{syn}}[h]_{i \bmod n_{\text{syn}}}$ when sizes differ, where n_{syn} is n). These are not attacks; they calibrate what a given feature’s R_{adv} score means, and under rarity weighting all three sit near 100/C for a C -valued feature. The one revealing exception is **copy** against RankSwap (49.3 in Table 1): RankSwap leaves most columns untouched, so copying the release returns the originals verbatim: a property of the RankSwap technique, not of any attack.

Reconstructing each feature in isolation. The dominant paradigm, due to Stadler et al. [58], treats the release (D_{syn}) as *training data*: for each hidden feature h , fit a supervised model with the QI columns as inputs and $D_{\text{syn}}[h]$ as the label, then apply it to the target’s QI. Any classifier serves; we evaluate logistic regression, naive Bayes, k -nearest neighbors, an SVM, random forests [9], gradient-boosted trees [37], a multilayer perceptron (MLP), and (new to this setting) **TabPFN** [32], a pre-trained transformer that produces a tabular predictor in a single forward pass. For a *continuous* hidden feature the same recipe substitutes a regressor for the classifier

⁴This setting is the adversarial mirror of *missing-data imputation*, used routinely in healthcare [16, 60]: MICE [66] is per-feature chained regression, MissForest [59] is our random-forest attack, and generative imputers such as GAIN [73] parallel our diffusion attacks. Only the intent differs (cooperatively filling missing values vs. adversarially inferring withheld ones), so advances in either literature transfer directly.

Table 1: Mean reconstruction accuracy (R_{adv} , %): each cell averages over all hidden features and 5 disjoint training sets, with rows grouped by taxonomy branch (Figure 2). The three reference points are not attacks, and the *Best ensemble/Best chain* rows give each enhancement’s strongest configuration (Appendix D.2). $\dagger$ = novel attack. (Adult, 10k rows, QI_{demo}.)

Attack	RankSwap	Cell Supp.	Synthpop	TVAE	CTGAN	ARF	TabDDPM	MST $\epsilon=0.1$	MST $\epsilon=1$	MST $\epsilon=10$	MST $\epsilon=100$	MST $\epsilon=1000$	AIM $\epsilon=1$	Avg.
<i>Reference points (not attacks)</i>														
Mode	10.5	10.5	10.5	10.5	10.5	10.5	10.5	9.5	10.5	10.3	10.3	10.3	10.5	10.4
Random	10.5	10.5	10.5	10.5	10.3	10.5	10.5	9.5	10.5	10.3	10.3	10.3	10.5	10.4
Copy	49.3	10.3	10.5	10.4	10.3	10.4	10.5	9.4	10.5	10.2	10.2	10.2	10.5	13.3
<i>Each feature in isolation: classifiers & column-marginal</i>														
KNN	25.5	55.7	26.9	22.1	19.3	18.6	39.3	13.1	15.4	13.2	16.3	16.3	15.4	22.9
Naive Bayes	22.2	46.3	31.1	18.6	18.4	18.8	31.5	13.0	15.9	14.1	16.2	16.2	17.1	21.5
Logistic Regression	15.6	17.3	16.7	16.1	15.3	14.7	16.9	10.9	12.0	12.3	12.8	13.3	12.6	14.3
SVM	21.5	26.0	25.5	21.1	20.6	19.5	26.1	11.7	14.2	11.7	15.1	14.8	14.4	18.6
Random Forest	25.7	52.3	28.8	24.1	21.2	20.5	38.9	13.1	20.9	19.2	18.9	18.5	20.9	24.8
LightGBM	23.7	34.2	27.9	23.1	22.6	22.7	23.8	13.5	21.8	18.3	18.2	17.8	21.1	22.2
MLP	23.3	33.9	30.1	27.0	24.3	22.9	34.8	13.9	21.5	19.1	19.2	19.4	23.8	24.1
TabPFN	23.0	46.3	27.6	23.6	21.2	19.7	34.7	12.7	15.2	13.6	15.9	15.2	14.7	21.8
Best ensemble	25.3	42.1	29.7	25.2	23.3	22.4	36.0	13.5	15.5	15.8	15.5	16.1	22.0	23.3
<i>Feature-correlated: autoregressive</i>														
ARFFormer $\dagger$	22.6	30.0	29.7	27.4	24.6	23.1	31.8	13.1	21.1	18.6	18.4	18.5	23.7	23.3
Best chain	25.5	54.3	28.1	23.3	20.0	19.8	39.2	13.3	15.0	15.4	15.6	15.9	20.0	23.5
<i>Feature-correlated: row-wise message passing</i>														
CoBP-RA $\dagger$	26.7	53.1	30.7	27.6	24.9	23.8	40.4	13.8	22.8	20.5	19.9	20.4	22.4	26.7
<i>Feature-correlated: joint generative conditioning</i>														
MultiHeadMLP $\dagger$	23.4	39.2	30.1	26.4	23.5	22.3	33.5	13.7	15.7	15.6	15.8	15.4	23.8	23.0
CondMST $\dagger$	20.9	27.9	27.4	23.0	21.4	18.2	27.1	14.8	20.0	17.6	18.0	18.1	22.1	21.3
CondDDPM $\dagger$	22.8	28.0	26.7	22.3	18.5	18.0	29.1	14.0	20.0	18.6	18.7	18.7	21.2	21.3
CondRePaint $\dagger$	14.8	18.2	18.5	15.9	13.4	12.8	19.3	11.7	14.2	12.9	12.7	12.5	15.1	14.8
Avg.	22.5	31.7	24.1	20.6	18.8	18.1	27.0	12.4	16.6	15.1	15.7	15.6	17.6	

(we use a random-forest regressor and Bayesian ridge regression). *Ensembling* belongs to the same branch: any set of these classifiers can be combined by voting or averaging. Among these per-feature classifiers the random forest is the strongest in our benchmark and the MLP a close second (Table 1).

A different approach replaces the classifier with a *column-marginal* model: the linear-program (LP) attack of Annamalai et al. [4] assigns each record a soft membership score so that the release’s implied counting-query answers are matched, then rounds to a prediction. The LP handles one low-cardinality column at a time and grows intractable as the value count rises; on single binary targets it does not beat a random forest or MLP, and it is infeasible for the high-cardinality, many-feature NIST CRC setting. We therefore report it only on binary targets (Section 3) and defer extended results to Appendix E.2.

Exploiting correlations among hidden features. Per-feature attacks discard a real signal: the hidden features are themselves correlated, and the release encodes those correlations. Three designs evaluated in this paper exploit them, and (apart from classifier chaining Read et al. [53]) each is a contribution of this paper.

Autoregressive attacks predict hidden features in sequence, feeding each prediction back into the next. *Chaining* [53] (see Appendix D.2) wraps any in-isolation attack; **ARFFormer** (*autoregressive feature transformer*) instead learns the dependencies directly, encoding the QI with multi-head self-attention and predicting the hidden features autoregressively. New to this setting, it tracks the MLP closely.

Row-wise message passing is our strongest design. **CoBP-RA** (*correction with belief propagation*, the “cobra” attack), not previously described in the literature, keeps the per-feature random-forest posteriors as unary beliefs but couples them on a graph: it takes the maximum-spanning-tree of pairwise mutual information over the hidden features, attaches to each edge a local pointwise-mutual-information table estimated from the synthetic neighbors nearest the target in QI space, and runs one pass of sum-product belief propagation. The local tables are the crux: because the random forest already conditions on the QI, a global co-occurrence table would double-count it, whereas a neighbor-local table isolates the *residual* dependence between two hidden features that survives that conditioning. Belief propagation then lets a confident feature correct an uncertain neighbor: if the synthetic records around a target make the joint value (B, C) far more common than (A, C) ,

the belief for the first feature shifts from A toward B , a correction no independent classifier can make. By default CoBP-RA also adds the QI values as observed graph nodes and damps messages from high-entropy senders, the best configuration we found and the one we report throughout. (A further *column-wise* pass, reconciling each feature’s aggregate prediction with its synthetic marginal between the row-wise passes, gave no consistent gain, so CoBP-RA stays purely row-wise.) For continuous data we discretize each hidden feature into equal-frequency bins of the synthetic column, run the same message passing on the bin labels, and decode each prediction back to its bin midpoint (Table 16).

Joint generative conditioning turns a generative model into the attack: instead of predicting features one by one, it samples all hidden attributes *at once* from an SDG model conditioned on the target’s QI, so the sample respects the full joint distribution. We explore this region of the attack taxonomy with four attacks not previously described in the literature. The deep-generative form includes **MultiHeadMLP** (a single network with one head per feature, predicting all hidden features jointly), **CondDDPM** (a tabular diffusion model whose reverse process is conditioned on the QI during training), and **RePaint/CondRePaint**, which port the image-inpainting RePaint scheme [40] to tables: at each denoising step the QI columns are overwritten with a forward-noised copy of their true values, anchoring the generated hidden columns to the observed QI with no change to the trained model [31]. The graphical-model form is **CondMST**: we fit an MST-style graphical model [45] on the release and condition it on the QI to sample the hidden features jointly (with bounded- and independent-clique variants). The shared *Cond-* prefix marks this conditioning recipe; CondMST and CondDDPM are named for the MST and TabDDPM generators they adapt, but the construction is general (each is run against *all nine* generators, not only its namesake), and the details appear in Appendices C.1 and C.2.

What the benchmark shows. Table 1 presents a first series of results, pitting every attack against all nine generators (MST at five privacy budgets) on the Adult dataset (the five datasets are summarized in Table 3 and described in Appendix A.1). Throughout the paper, every reported score is averaged over 5 disjoint training samples. Two patterns dominate, and both point the same way.

First, *the generator matters far more than the attack*. Down any column, the best and worst attacks differ by roughly 8–20 points; across any row the same attack swings much more: a random forest ranges from 18.5 against near-noiseless MST to 52.3 against CellSuppression. The vertical **Avg.** row makes the ordering plain: de-identification (CellSuppression 31.7, RankSwap 22.5) and the highest-fidelity generator (TabDDPM 27.0) are the exposed regimes, while DP marginal synthesis collapses toward the Mode reference as the budget tightens (MST $\epsilon=0.1$: 12.4). Section 3 dissects this axis and why it, not the attack, is decisive.

Second, *no attack wins everywhere, and sophistication buys little*. KNN leads on CellSuppression, where retained rows are real records and the attack is effectively a lookup; naive Bayes is the best on Synthpop, whose sequential generation mirrors its column-wise independence; on TVAE, CoBP-RA, ARFFormer, and MLP cluster at the top (27.6, 27.4, 27.0), separated by under a point. **CoBP-RA is the best attack on average** (26.7) and is best or near-best in

almost every column, but its margin over a plain random forest (24.8) is under two points, and its belief-propagation correction helps most exactly where joint structure survives (the high-fidelity generators) and least against DP noise. The joint-generative attacks are a cautionary case: CondDDPM and CondRePaint trail simple classifiers even on TabDDPM-generated data, because 10k synthetic rows are too few to pin down their parameters.

Can attack combinations increase success rates? Two leaves of the taxonomy combine attacks rather than introduce new ones: *ensembling* (voting across in-isolation classifiers) and *chaining* (autoregressive prediction). Both helped our NIST CRC submission, yet in controlled experiments neither reliably beats the best single attack (our strongest ensemble and strongest chain both land below CoBP-RA alone), and in each case an oracle analysis shows the ceiling is structural, not algorithmic. For *ensembling*, the attacks really are complementary: an oracle selecting the correct attack per record would gain over eleven points. But that gap cannot be closed. Nothing in the QI marks which records a given attack will get right, per-feature selection recovers almost none of the gain, and a shared core of hard records defeats every attack at once, so even tuning the ensemble weights directly on the test labels yields no improvement over the single best attack. *Chaining* fails for the opposite reason: an oracle fed the *true* earlier hidden values gains only a few points (+7.2 for a random forest, +1.5 for an MLP), so error propagation is *not* the bottleneck; the synthetic data does not preserve the higher-order conditional structure autoregression would exploit, even for a generator as faithful as TabDDPM. We develop both negative results, and the oracle analyses that localize their ceilings, in Appendix D.2.

Takeaway

On a fixed release, switching *attacks* moves risk by a few points; switching *generators*, by tens. CoBP-RA is our strongest attack, yet the more striking pattern is how close simple methods come: a plain random forest lands within a couple of points of it, and the best attack is often mechanism-specific (naive Bayes rivals far richer models on column-sequential Synthpop). Sharper attacks may yet emerge, but on every release, the generator, not the adversary, drives exposure.

3 How Does the Scenario Shape Risk?

Section 2 fixed almost everything but the attack; here we do the reverse, holding the attack family roughly fixed and sweeping the *scenario* in three groups: the generator and its privacy budget (Section 3.1), the data domain, rows, and features (Section 3.2), and the adversary’s quasi-identifier (Section 3.3). One ordering recurs: the release governs how much can be reconstructed, the adversary merely realizes it; the interesting structure is in the exceptions, and in how each dimension moves the ceiling.

3.1 SDG Method and Privacy Budget

As stated in Section 2, the SDG method is the single largest lever on reconstruction risk, and the ordering follows how much each

Table 2: Synthetic-data quality profile and reconstruction vulnerability. Metrics (arrows $\uparrow/\downarrow$ in the column headers give the favorable direction): TSTR ratio, Train-on-Synthetic / Test-on-Real F1-macro (>1 = synth beats the real-data baseline); Mean JSD, Jensen–Shannon divergence over marginals; Col. shapes, Col. pairs, SDV marginal and bivariate fidelity (1 = indistinguishable from real); Pairwise TVD, mean TV distance over pairwise marginals; Corr. diff, mean absolute error between real and synthetic correlation matrices; Wass. OHE [28], a marginal-fidelity score: each feature is one-hot encoded (categorical) or equal-depth binned (continuous), and the per-feature L_1 distances between the real and synthetic marginal distributions are summed; RF R_{adv} , a Random Forest attack’s R_{adv} . *Train–Train* (real vs. real) is a fidelity benchmark; best per column excluding it in bold. Per-dataset breakdowns in Appendix E.11. (Adult, 10k rows.)

SDG method	TSTR ratio $\uparrow$	Mean JSD $\downarrow$	Col. shapes $\uparrow$	Pairwise TVD $\downarrow$	Col. pairs $\uparrow$	Corr. diff $\downarrow$	Wass. OHE $\downarrow$	RF R_{adv}
MST ($\epsilon=0.1$)	0.616	0.101	0.746	0.154	0.754	0.074	4.326	13.1
MST ($\epsilon=1$)	0.628	0.033	0.908	0.064	0.857	0.072	2.735	20.9
MST ($\epsilon=10$)	0.769	0.008	0.790	0.039	0.809	0.054	2.521	19.2
MST ($\epsilon=1000$)	0.750	0.002	0.785	0.037	0.835	0.053	2.501	18.5
AIM ($\epsilon=1$)	0.949	0.045	0.917	0.058	0.871	0.042	4.731	20.9
RankSwap	0.972	0.000	0.987	0.000	0.929	0.012	0.097	25.7
Cell Supp.	1.021	0.016	0.993	0.015	0.990	0.005	0.216	52.3
Synthpop	1.002	0.011	0.993	0.024	0.982	0.011	0.272	28.8
TVAE	0.982	0.103	0.871	0.139	0.839	0.045	3.066	24.1
CTGAN	0.964	0.089	0.862	0.168	0.843	0.035	2.727	21.2
ARF	0.973	0.092	0.894	0.187	0.810	0.014	2.673	20.5
TabDDPM	1.018	0.016	0.990	0.022	0.982	0.014	0.331	38.9
<i>Train–Train</i>	—	0.014	0.990	0.026	0.980	0.010	—	—

Table 3: Datasets used in this study. Feature counts are categorical/continuous; ‘✓’ marks data of a sensitive nature. *NIST Arizona is IPUMS 1940 census microdata from the NIST Privacy CRC; we use its 25-feature setting.

Dataset	Feat.	Rows	Domain	Sens.	Ref.
Adult	15 (9/6)	48k	Demographic	✓	[7]
NIST Arizona*	25 (19/6)	235k	Demographic	✓	[50]
California Housing	9 (0/9)	22k	Housing		[56]
NIST SBO	130 (125/5)	1.6M	Finance	✓	[51]
CDC Diabetes	22 (19/3)	200k	Healthcare	✓	[13]

Table 4: Reconstruction accuracy (R_{adv} , %) across the full MST privacy-budget sweep, for the QI_{large} and $QI_{behavioral}$ sets (defined in Table 23). Each row is nearly flat (above $\epsilon \approx 1$ the worst-case DP guarantee barely tracks empirical risk), and the residual within-row variation is sample noise, not an ϵ -trend. Memorization deltas were all $|\Delta| < 1$ pp (see Section 4.1). $\dagger$ = novel attack. (Adult, 10k rows.)

Attack	$\epsilon =$	0.1	0.3	1	3	10	30	100	300	1000
<i>QI_{large} (10 demographic features)</i>										
Random Forest	11.8	10.6	11.9	11.7	11.2	12.6	12.5	13.6	12.9	
Naive Bayes	12.7	10.1	10.0	10.1	10.0	10.1	10.1	10.0	10.1	
CoBP-RA †	12.9	12.9	14.5	14.7	13.5	14.2	14.7	14.6	14.6	
<i>$QI_{behavioral}$ (6 behavioral/financial features)</i>										
Random Forest	20.8	21.3	21.7	21.3	22.1	22.2	22.8	22.0	22.7	
Naive Bayes	19.3	19.0	19.2	19.2	19.2	19.2	19.2	19.1	19.2	
CoBP-RA †	21.7	22.5	23.5	24.5	24.3	24.5	24.4	24.1	24.7	

method transforms the data: the *least* transformed releases leak the most. De-identification leads: cell suppression enforces k -anonymity [62] by dropping records whose QI combination is too rare and releasing the rest unmodified, so KNN’s 55.7% on Adult

(Table 1, QI_{demo} defined in Section 3.3) is little more than a lookup against surviving real records; rank swapping permutes values within blocks, preserving marginals while leaving enough joint structure for a random forest to reach 25.7%.

In general, as seen in Table 2, vulnerability tracks quality of the generated synthetic data. Across all nine generators, downstream utility (measured as the F1-score of a classifier trained on the synthetic data and evaluated on real data, TSTR) and RF reconstruction are rank-correlated at $r_s=0.80$ ($p=0.01$). The coupling holds *within* an SDG mechanism too: at a matched $\epsilon=1$, AIM buys far more utility than MST (TSTR 0.949 vs. 0.628; Table 2) and, following the same logic, leaks slightly more: its mean reconstruction across attacks edges above MST’s at that budget. Fidelity and exposure are two readings of one dial.

But it is *joint* fidelity that tracks risk, not single-feature marginal fidelity. The coupling noted above is with downstream *utility* (TSTR), which reflects joint, conditional structure; how closely a release reproduces the real *single-feature* marginals (mean JSD) shows no such relationship: its Spearman correlation with reconstruction is under 0.25 in magnitude for every attack we test. TVAE and MST at $\epsilon=0.1$ make the point: their marginals diverge about equally (mean JSD 0.103 vs. 0.101), yet TVAE is reconstructed nearly twice as accurately (24.1% vs. 13.1%), tracking joint fidelity (TSTR 0.982 vs. 0.616) instead; conversely, near-noiseless MST at $\epsilon=1000$ has near-perfect marginals (JSD 0.002) yet *lower* risk (18.5%) than deep generators with marginals an order of magnitude worse. A data holder cannot read privacy off noised marginals: their distance from the real ones says little about how reconstructable the records are (Table 2; per-dataset breakdowns in Appendix E.11, Tables 19–22).⁵

⁵A preliminary analysis does not find the feature-correlation attacks (e.g. CoBP-RA, MultiHeadMLP) gaining over feature-isolated ones *specifically* on higher-joint-fidelity

Table 5: Reconstruction accuracy on binary hidden features, averaged over all SDG methods. R_{adv}^{tr} = accuracy on training-set targets; $\Delta = R_{adv}^{tr} - R_{adv}^{NT}$ is the gap to held-out (non-training, NT) targets from the same population, the individual-level (memorization) leakage (Section 4.1). Best R_{adv}^{tr} per column in bold; the shaded row is the LP attack proposed as state-of-the-art by Annamalai et al. [4]. The lower block adds two attacks new to this setting, TabPFN [32] and CoBP-RA[†]; neither wins a column in this single-hidden-feature binary setting (where inter-feature propagation is moot), though CoBP-RA stays within a point of the best in every column. [†] = novel attack.

Attack	Adult 1k				Adult 10k				Arizona 10k		CDC 1k					
	income		sex		income		sex		sex		Diabetes		HighBP		Stroke	
	R_{adv}^{tr}	Δ	R_{adv}^{tr}	Δ	R_{adv}^{tr}	Δ	R_{adv}^{tr}	Δ	R_{adv}^{tr}	Δ	R_{adv}^{tr}	Δ	R_{adv}^{tr}	Δ	R_{adv}^{tr}	Δ
Random	50.0	+0.0	50.0	+0.0	50.0	+0.0	50.0	+0.0	50.0	+0.0	50.0	+0.0	50.0	+0.0	50.0	+0.0
KNN	63.7	+5.3	67.3	+5.7	69.5	+4.6	75.2	+4.0	66.6	+4.2	60.7	+7.1	64.3	+7.4	59.1	+7.5
Random Forest	67.0	+5.5	75.1	+3.9	69.6	+4.1	80.4	+3.3	70.8	+3.8	58.4	+6.4	68.1	+6.2	56.0	+5.7
MLP	70.1	+2.5	75.1	+2.8	71.8	+2.8	80.6	+1.3	71.1	+1.3	60.2	+5.0	65.2	+2.9	57.5	+4.1
TabPFN	60.8	+3.3	71.2	+2.4	61.8	+0.2	73.4	+0.0	69.9	+0.1	53.8	+2.7	64.3	+2.9	51.8	+1.6
CoBP-RA [†]	67.8	+5.2	75.0	+4.3	70.5	+3.8	80.2	+3.4	70.8	+3.9	58.2	+6.2	67.2	+5.5	55.7	+5.4
Linear Program	63.3	+4.0	66.1	+3.2	67.9	+2.4	76.1	+2.3	63.8	+1.2	59.4	+6.1	63.0	+5.5	58.8	+7.3
Avg.	63.1	+3.3	68.1	+2.8	65.8	+2.3	73.8	+1.7	65.8	+1.8	57.4	+4.2	63.0	+3.9	55.6	+4.1

This is why differential privacy helps mainly below $\epsilon=1$. For MST on Adult, RF reconstruction drops sharply at $\epsilon=0.1$ (to 13.1%); from $\epsilon=1$ to $\epsilon=1000$ it barely stirs (20.9% $\rightarrow$ 18.5%; Table 1), and the full nine-budget sweep is a flat plateau for both RF and Naive Bayes (Table 4), even as Wasserstein distance keeps falling toward the deep generators. Above $\epsilon \approx 1$ the binding constraint is thus MST’s pairwise *capacity*, not its noise. This is a clean qualitative break from *membership* inference, which on these same synthesizers keeps strengthening with ϵ out to 1000 [28]: reconstruction is capacity-capped where membership is not. Two edges qualify the plateau. Its onset is adversary-dependent: a richer QI saturates the ceiling even under heavy noise, i.e., small ϵ (RF on QI_{behavioral} spans just 20.8 $\rightarrow$ 22.8% across the sweep), so tight budgets protect weak adversaries most. And $\epsilon=0.1$ is anomalous in both fidelity and attackability: the marginals are so noised that *every* attack, baselines included, dips below its value at all higher budgets, as even easy low-cardinality features wash out (education-num and hours-per-week fall to 0%, recovering only at $\epsilon=1$; Appendix Table 13). This baseline dip is general, not an Adult artifact: the noised marginals shift the values the reference points read off, so it recurs on every DP dataset (CDC’s Mode falls to 41.7% at MST $\epsilon=0.1$, Table 11).

A flat *measured* curve is not formal safety, however: at large ϵ the worst-case guarantee is nearly vacuous. Randomized response is ϵ -DP yet preserves over 90% of a sensitive bit once $\epsilon \geq \ln 9 \approx 2.2$ (*blatant non-privacy* [18]), and real deployments run far higher still, the 2020 US Census at $\epsilon \approx 19.61$ [2].

Inside the plateau. Table 4 traces the full sweep for two quasi-identifiers. The aggregate flatness hides orderly per-feature motion (Appendix E.3): most features switch on at $\epsilon=1$ and then freeze, while occupation (the highest-cardinality target) is the lone feature to climb monotonically across the range, since its conditional structure needs the most marginal budget to resolve.

releases: their edge over a per-feature random forest tracks joint fidelity no more than the isolated attacks’ does. This bears on that coupling alone, not on overall strength.

3.2 Domain, Features, and Scale

Raw R_{adv} is not comparable across our five datasets (Table 3): both the baseline and the cardinality ceiling shift with the data. Two structural drivers explain most of the per-feature variation. The first is **cardinality**: on Adult, high-cardinality targets sit near the floor (fmlwgt, $C=8,477$, scores 0.8%; capital-gain, $C=98$, 7.0%) while low-cardinality ones are exposed (income, $C=2$, 77.4%). The second, and more telling, is **predictability from the QI**: education-num ($C=16$) scores 89.8% (above even binary income) because it is essentially a re-encoding of QI education, whereas income is genuinely harder to guess from demographics. Once cardinality is low, what the QI *determines* sets the score. The baseline moves for the same reason: CDC Diabetes’ mostly-binary features yield a Mode baseline of 42.5% (majority-class guessing earns $\approx 50\%$ under rarity weighting), against Adult’s 10.5%, where high-cardinality columns earn Mode almost nothing. Aggregate scores must therefore be read with the per-feature view: reconstructing income and occupation is real harm even when fmlwgt is unrecoverable.

Across domains. The ceiling, not the attack, is what shifts across datasets (QI memberships in Appendix E.8). On Adult the best attack averages 21.7% across generators, 11 pp above Mode; the other four datasets each add a distinct lesson.

CDC Diabetes (full per-generator results in Appendix E.4, Table 11) is the most exposed categorical dataset under de-identification (KNN hits 83.5% on rank swapping), and TabDDPM, at 75.7%, rivals it despite being fully synthetic, since diffusion captures binary health dependencies cleanly even from 1k rows. DP is correspondingly tight: across the entire MST sweep KNN moves only 41.5 $\rightarrow$ 44.2%, because binary marginals already carry most of the signal and leave little residual structure. CoBP-RA leads or ties the field on most CDC and Arizona generators, extending to these datasets the best-or-near-best behavior that the Adult benchmark established in Section 2.

Feature count acts through heterogeneity, not the average. On NIST SBO (130 mostly-binary financial features) the aggregate gap above Mode is a modest 5 pp (Appendix Table 10), but it splits into two tiers: a handful of employee-benefit flags (health insurance, retirement, holidays) leak 12–14 pp because they are predictable from firm sector and size, while noisy high-cardinality financials (receipts, payroll) leak under 2 pp: the *fnlwtg* pattern at scale. Only 5 of 123 hidden features clear 10 pp, but they are among the most sensitive business facts. NIST Arizona (25 census features) behaves much like Adult, with the best attack at 22.7%.

Continuous data (California Housing, scored by NRMSE) shifts the quality landscape, not the conclusion. Attacks still beat mean-imputation (RF NRMSE 0.038–0.112 against a 0.09–0.18 baseline), but the trade-off sharpens into a cliff: MST and AIM at $\epsilon \leq 1$ discretize the columns so aggressively that downstream utility collapses to TSTR = 0.000 (Appendix Table 22), recovering only by $\epsilon=1000$, whereas non-DP Synthpop and TabDDPM deliver real utility at real reconstruction risk. For continuous data, no budget buys both.

Scale cuts both ways. The effect of dataset size is not monotone: it is set by the mechanism (Table 12). Growing CDC Diabetes from 1k to 100k makes *de-identification* markedly more exposed (RF on cell suppression 58.6 $\rightarrow$ 87.1%, as more QI groups clear the k -anonymity threshold and ship verbatim) yet makes *diffusion* safer (TabDDPM 75.1 $\rightarrow$ 47.7%, as abundant data dilutes per-record memorization); Adult shows the same diffusion effect (48.6 $\rightarrow$ 38.9%, 1k $\rightarrow$ 10k), while MST inches up as its marginals sharpen. Attacker-side scale interacts with this: TabPFN, a prior-fitted transformer built for small tables, matches the best classifiers at 1k (on Adult it ties RF on Cell Supp., 51.1 vs. 49.6, and on TabDDPM, 48.9 vs. 48.6) but falls behind by 10k (Cell Supp. 46.3 vs. 52.3; MST $\epsilon=1$ 15.2 vs. 20.9), where its small-sample prior stops paying off. More data is not uniformly more or less private; it depends on what consumes it.

3.3 Quasi-identifying Features

The quasi-identifier (the features the adversary is assumed to know) is the last design axis. Our default QI_{demo} is a set of standard demographic attributes per dataset (for Adult: age, sex, race, native-country, education, marital-status; full per-dataset definitions in Appendix E.8, Tables 23–27); such auxiliary knowledge is empirically cheap, as Sweeney [61] re-identified 87% of Americans from {ZIP, birth date, sex} and Rocher et al. [54] estimate 99.98% from 15 attributes. Table 6 varies QI *size* and, at a matched size, QI *composition*, each scored over a fixed hidden-feature intersection for comparability (a parallel CDC analysis is in Appendix E.6).

Composition beats size. Adding features helps only until the QI captures the main predictors: on Adult, RF climbs +12.0 pp from QI_{tiny} (3 features) to QI_{demo} (6), then just +3.4 pp more to QI_{large} (10). Which features are known matters more. At a matched $|QI|=6$, a behavioral QI (employment, hours) gives RF 28.3% versus 24.7% for the demographic one: a +3.6 pp swing from composition alone, larger than the gain from four extra features and stable across samples and generators. This mirrors Narayanan and Shmatikov [49]: behavioral side-channels re-identify where demographics cannot.

More can be less. Composition can even invert the effect of size. On CDC Diabetes (Appendix E.6), a 10-feature *behavioral* QI

scores *below* the same-size demographic one across every attack (RF 33.3 $\rightarrow$ 29.2%): the behavioral set absorbs the lifestyle features that were the easy hidden targets, leaving a harder residual. The adversary knows more, but the wrong things. And the per-feature caution returns in force: a favorable *average* can still hide near-perfect reconstruction of a single sensitive outcome (Diabetes, Stroke), so QI design must be judged feature by feature, not in aggregate.

Takeaway

Risk is set by the release, then shaped at the margins: by the *data's* cardinality and QI-predictability, by *scale* (which exposes de-identification yet dilutes a diffusion model's memorization), and by the *composition*, not the size, of what the adversary knows. The sharpest surprise is the shape of differential privacy. For MST (the one synthesizer whose budget we sweep), protection accrues below $\epsilon \approx 1$ and then *plateaus*: above it, reconstruction is capped by the marginal model's *capacity*, not its noise, so risk barely moves out to $\epsilon=1000$. This is in pointed contrast to membership inference, which on the same synthesizer keeps leaking as ϵ grows.

4 What Does “Success” Actually Mean?

The previous two sections measured *how much* can be reconstructed and *under what conditions*. But a single number, R_{adv} , conflates phenomena that carry very different privacy meaning. A record can be reconstructed because the generator memorized it, or because the release merely preserves population structure that would predict the same value for anyone like them. This section pulls those apart. We ask three questions of every success: is it memorization or inference (§4.1)? Does the signal that drives reconstruction also reveal *membership* (§4.2)? And is the risk it measures shared evenly across people, or concentrated on a few (§4.3)? The answers recast the rankings of the previous sections: the generator that leaks most on average is not the one that most endangers individuals.

4.1 Memorization or Generalization?

Two mechanisms produce a correct reconstruction. A generator may *memorize* an individual training record, letting the attacker read it back; or the release may faithfully preserve *distributional* structure, letting the attacker *infer* a hidden value that any similar person would share. The distinction is the crux of what “leakage” means: distributional inference is intrinsic to any useful release (it is the same signal that makes synthetic data analytically valuable), whereas memorization is a failure of the privacy mechanism, exposing a specific person who happened to be in the data.

Cohen et al. [15] formalize exactly this boundary as *narcissus resiliency*: a reconstruction counts as genuine only if it could not have been produced by a process that never saw the target; otherwise the attacker is, fittingly for the name, merely *admiring a reflection* of the population rather than recovering the individual. Our *memorization test* is an empirical proxy motivated by it. For each generator we run the attack against training targets ($R_{\text{adv}}^{\text{tr}}$) and against disjoint *non-training* targets from the same distribution ($R_{\text{adv}}^{\text{NT}}$), the oblivious baseline. The gap $\Delta = R_{\text{adv}}^{\text{tr}} - R_{\text{adv}}^{\text{NT}}$ is the *memorization gap*, the departure from narcissus resiliency: $\Delta \approx 0$

Table 6: Reconstruction accuracy by quasi-identifier setting. Within each section, cells average over the *intersection* of hidden features across that section’s QI variants, so columns are comparable within a section (per-variant QI memberships in Table 23); bold = row maximum within a section. $\Delta_{\text{size}} = \text{QI}_{\text{large}} - \text{QI}_{\text{tiny}}$, $\Delta_{\text{comp}} = \text{QI}_{\text{beh.}} - \text{QI}_{\text{demo}}$, and $\Delta_{t \rightarrow d}^*$ is the gain from adding 3 demographic features, scored on the broader 9-feature two-way intersection. (Adult, 10k rows, mean over all SDG methods.)

Attack	QI size sweep ($\nearrow$ adversary knowledge)					Composition (matched QI)		
	QI _{tiny} (3 kn.)	QI _{demo} (6 kn.)	$\Delta_{t \rightarrow d}^*$	QI _{large} (10 kn.)	Δ_{size}	QI _{demo} (6 kn.)	QI _{beh.} (6 kn.)	Δ_{comp}
Random	11.44	11.43	−0.00	11.45	+0.01	17.13	17.15	+0.02
KNN	13.00	23.56	+9.91	25.18	+12.17	25.10	28.50	+3.40
NB	13.80	23.03	+7.44	24.09	+10.30	24.79	21.35	−3.44
RF	12.48	25.31	+12.01	28.68	+16.20	24.73	28.34	+3.61
LGBM	12.39	21.87	+9.36	22.11	+9.73	22.41	22.30	−0.11
MLP	12.52	25.64	+11.02	26.56	+14.04	24.00	27.91	+3.91

Table 7: Memorization test: training vs. non-training accuracy (R_{adv}^{tr} , R_{adv}^{NT}) and gap Δ , alongside a simplified El Emam disclosure score (d_S , higher = more risk) [23]. Bold Δ marks the three generators that genuinely memorize individuals. (Binary feature income, Adult, 1k rows, RF+KNN+MLP averaged.)

SDG method	R_{adv}^{tr}	R_{adv}^{NT}	Δ	d_S
Cell Supp.	92.7	72.5	+20.1	0.417
RankSwap	81.8	68.3	+13.5	0.089
TabDDPM	86.1	72.8	+13.4	0.281
Synthpop	73.4	71.8	+1.6	0.148
CTGAN	50.5	49.5	+1.0	0.055
AIM ($\epsilon=1$)	57.5	56.5	+1.0	0.019
MST ($\epsilon=1$)	53.7	53.9	−0.3	0.018
MST ($\epsilon=10$)	61.3	60.6	+0.7	0.011
MST ($\epsilon=1000$)	58.1	58.7	−0.6	0.010
TVAE	63.1	62.8	+0.3	0.104
ARF	65.1	64.8	+0.3	0.045

means every success would have occurred without the target in the data (pure distributional inference), while a large Δ isolates genuine individual memorization. Table 7 reports it for the binary income feature (RF, KNN, MLP averaged), alongside the classical disclosure score discussed below.

The generators split into three regimes. **De-identification leaks individuals.** CellSuppression and RankSwap carry the largest gaps ($\Delta = +13$ to $+20$ pp): they edit rather than resynthesize, so original rows survive and stay individually traceable: the high de-identification scores of Section 2 are, in substantial part, memorization rather than inference. **Most generative and DP methods do not.** Synthpop, CTGAN, TVAE, ARF, AIM, and MST all sit at $|\Delta| \leq 1.6$ pp; whatever their magnitude, their reconstruction success is distributional. The slightly negative MST gaps (−0.3, −0.6) are within sampling noise and confirm the DP guarantee empirically; tellingly, they do not grow from $\epsilon=1$ to $\epsilon=1000$, the same insensitivity to the budget we saw for aggregate risk in Section 3.

Diffusion is the exception that proves the rule. TabDDPM’s gap ($\Delta = +13.4$ pp) is an order of magnitude above every other

resynthesizing generator (Synthpop +1.6, CTGAN +1.0, TVAE/ARF +0.3) and rivals RankSwap, which copies records outright: a *fully synthetic* generator is memorizing individuals. The mechanism mirrors what Carlini et al. [11, 12] document for language and image diffusion models: high-capacity generators interpolate between generalization and memorization, and for records with rare QI combinations they fall back on encoding the individual. The per-feature view localizes it: the gap concentrates on identity-bearing attributes (occupation, workclass: $\Delta \approx +3$ –4 pp) rather than predictable ones (income, education-num: +1 pp). This sharpens the Section 3 finding that fidelity and exposure rise together: for TabDDPM the *same* capacity that earns it the top aggregate risk in Table 1 is what makes it memorize atypical people. The continuous domain agrees: on California Housing the three-regime structure reappears with the sign reversed (TabDDPM RF $\Delta = -0.034$, MST $|\Delta| \leq 0.001$; Appendix E.9, Table 16), so the signature is structural, not an artifact of categorical scoring.

Classical disclosure risk. The rightmost column of Table 7 places our adversary against the bottom rung of the classical ladder: a simplified form of the El Emam *et al.* identity-disclosure-risk model [23], $d_S = \frac{1}{n} \sum_s E_s I_s R_s$ (the subscript S labels the synthetic release; the indicators are per record s), which credits a match only when a record sits in a QI equivalence class of size $\leq k$ (E_s), its sensitive value is correctly inferred (I_s), and that value is rare (R_s): a lookup attacker with no learning (we drop their error/verification adjustment to use the indicator core as a pure matching baseline). d_S reproduces the same ordering as R_{adv} (CellSuppression most exposed, DP least), which cross-validates the ranking, but it badly understates magnitude: it scores CellSuppression at 0.42 where a KNN attacker reconstructs 64%, because ML attacks generalize *across* equivalence classes rather than requiring exact QI matches. Its one independent signal echoes the memorization finding: TabDDPM’s d_S (0.28) exceeds every other resynthesizer, reached with no model at all.

4.2 Relationship to Membership Inference

Membership inference (MIA) asks a strictly easier question than reconstruction (was this record in the training set?), yet it is the standard currency of privacy auditing [10, 57], including on synthetic data [28, 30, 65]. The two have been scored on incomparable scales, which has kept them apart. We bridge them with a *reduction* in the complexity-theoretic sense: a procedure that solves membership inference by calling reconstruction as a subroutine. Given a candidate record, the adversary runs the full reconstruction attack on it and returns *how well it reconstructs* as the membership score, on the hypothesis that a generator encodes members’ conditional structure more faithfully. The reduction is tight in the resource that matters: it consumes nothing a MIA adversary lacks (only the synthetic data and the candidate’s QI), so any success it achieves is a success MIA could have claimed. If reconstruction is the stronger primitive, this reduction should make it a competent membership oracle, and we can ask whether the two paradigms flag the same *releases* and the same *people*. We compare it against two standard distance baselines [29, 58]: *SynthDistance* (negative distance to the nearest synthetic row) and *NNDR* (that distance taken relative to

Table 8: AUC of MIA (SynthDistance, NNDR) vs. RA-as-MIA across three datasets. RA-as-MIA uses RF R_{adv} as the membership signal.

	Cell-Supp.	RankSwap	TabDDPM	Synthpop	MST ($\epsilon=1$)	MST ($\epsilon=1000$)
Adult ($n = 10,000$, 15 features, QI_{demo})						
SynthDistance	0.96	0.82	0.77	0.55	0.50	0.50
NNDR	0.97	0.87	0.80	0.58	0.50	0.50
RA-as-MIA	0.65	0.57	0.62	0.52	0.49	0.50
CDC Diabetes ($n = 1,000$, 22 features, QI_{demo})						
SynthDistance	0.62	0.99	0.80	0.57	0.51	0.52
NNDR	0.64	1.00	0.81	0.60	0.51	0.53
RA-as-MIA	0.60	0.77	0.71	0.52	0.51	0.51
NIST Arizona ($n = 10,000$, 25 features, QI_{medium})						
SynthDistance	0.93	0.84	0.72	0.51	0.50	0.50
NNDR	0.96	0.91	0.80	0.56	0.52	0.52
RA-as-MIA	0.62	0.55	0.57	0.51	0.50	0.50

the nearest leave-one-out training neighbor). Table 8 spans the full vulnerability spectrum across three datasets.

Two things hold across all three datasets. First, every paradigm draws the same map of exposure: de-identification is wide open (CellSuppression and RankSwap NNDR up to 1.00), TabDDPM is meaningfully but less exposed, and both MST budgets sit at chance (≈ 0.5) irrespective of ϵ . Second (and this is the point of the reduction), RA-as-MIA traces that same map using reconstruction alone: where dedicated MIA succeeds it succeeds, and where MIA fails (MST) it fails too. Reconstruction success and membership read the same leakage, which is why the reduction works at all.

What it does *not* read is the same *individuals*. The agreement is aggregate, not per-record: for TabDDPM training records, RA-as-MIA and SynthDistance correlate at only Spearman $r=0.41$, and of 10,000 members, 3,269 are flagged *discordantly*: ranked above the median by one attack and below it by the other. The two attacks succeed on overlapping but distinct subpopulations, so an audit that runs only one will systematically miss the people the other would catch; neither subsumes the other. RA-as-MIA’s discriminative power is itself concentrated on outliers (AUC 0.74 on atypical records vs. 0.60 overall), the same concentration we turn to next.

4.3 Disparate Impact: Who Is Most at Risk?

A mean hides its tail. Risk concentrates on atypical records (Meeus et al. [46] call them “Achilles’ heels” in synthetic data, and Yaghini et al. [71] show disparate membership-inference vulnerability across demographic groups), and we ask how that concentration behaves across the reconstruction spectrum. We score each individual by feature-averaged row-level R_{adv} and label as *outliers* the 10% whose quasi-identifier profile is most atypical: those an Isolation Forest, fit on the public QI columns alone, most readily isolates from the rest of the population. Outlier status is therefore a property of how unusual a person looks to the adversary *before* any attack, not of the hidden values or whether reconstruction succeeds. The **outlier penalty** (the “Outlier $\times$ ” column) is then the mean row-level R_{adv} on the outlier 10% divided by the mean on the typical

90%: a value near 1 means reconstruction falls evenly, while a value above 1 means the attack reconstructs unusual individuals more accurately than typical ones. The remaining columns break the same score down by race and sex. Table 9 reports these, with one scale caveat: row-level R_{adv} weights a single person’s values by rarity, so most rows score far below the population aggregates of Section 2; compare *within* the table, not against Table 1.

The disparity is real, large, and (crucially) *opposite in direction* depending on the mechanism, which is why we report it per generator rather than per attack: the protection mechanism, not the attacker’s algorithm, decides who is exposed. Under high-fidelity synthesis, rarity is a liability. TabDDPM reconstructs outliers 3.4 $\times$ as well as typical records, and Amer.-Indian/Alaska Native individuals (1.1% of the data) at 7 $\times$ the rate of the White majority (8.4% vs. 1.2%): few synthetic records share a minority QI pattern, so each maps back to a single real person. Suppression inverts the sign: CellSuppression *removes* the most heavily suppressed minority records (perfectly protecting AI/AN individuals, 4.1%) while leaving the majority and partially-retained groups exposed (Black 61.0%, API 52.7%). Differential privacy alone spreads risk evenly: both MST variants stay within 1.5 $\times$ across every subgroup, the equity its guarantee promises.

Two checks confirm this concentration is a property of the *release*, not of our attacker. It is **attack-invariant**: from simple in-isolation classifiers to our strongest correlation-aware attack, CoBP-RA, the outlier penalty is fixed by the generator, not the algorithm: holding near $\times 2.2$ for CellSuppression and $\times 2.9$ for TabDDPM, while none lifts Synthpop or any DP release above $\times 1$. A stronger adversary recovers more on *average* but does not change *who* is exposed. The penalty also concentrates by a **property of the feature**: it falls on attributes with many rare values, where an outlier’s own value is itself identifying, and vanishes on those a single common value dominates, easy to guess for everyone but the outlier. Per-attack and per-feature breakdowns are in Appendix E.10 (Tables 17 and 18).

The lesson is not that one mechanism is fairer but that *mean R_{adv} is silent on this axis*: synthesis endangers the rare, suppression the common, and only per-subgroup, per-feature reporting tells them apart.

Takeaway

A single R_{adv} hides three different things. Most reconstruction is not a breach but *distributional inference* (true of anyone like the target), with genuine memorization the exception, confined to record-retaining de-identification and, alone among resynthesizers, high-capacity diffusion; and reconstruction tracks *membership inference* (the standard audit signal) closely enough to match it as a privacy measure while revealing more. Most striking, one mean can endanger *opposite* individuals: synthesis concentrates risk on the rare, suppression on the common. An aggregate score is where an audit starts, never where it ends.

5 Implications for Auditing and Policy

Across three chapters one ordering held: the release governs how much can be reconstructed, and the attacker only realizes it. For a

Table 9: Disparate impact. Mean and the per-race/per-sex columns are mean row-level R_{adv} (%); Outlier $\times$ is the *unitless* ratio of that score on the outlier 10% (the most QI-atypical individuals; Isolation Forest on the public QI, see text) to the typical 90% (>1 = outliers more exposed). The ‘%’ beside each subgroup header is its population share. AI/AN = Amer.-Indian/Alaska Native, API = Asian-Pacific Islander. Per-attack and per-feature breakdowns in Appendix E.10. (RF, Adult 10k, QI_{demo})

SDG	Mean (%)	Outlier $\times$	Mean row R_{adv} by race (%)					By sex (%)	
			AI/AN (1.1%)	API (3.1%)	Black (9.9%)	White (85.3%)	Other (0.7%)	Female (32.1%)	Male (67.9%)
Cell Supp.	33.1	$\times 2.3$	4.1	52.7	61.0	29.8	0.4	42.2	28.8
TabDDPM	1.4	$\times 3.4$	8.4	3.9	1.9	1.2	0.9	1.7	1.3
Synthpop	1.6	$\times 1.0$	2.2	1.2	1.2	1.7	0.4	1.5	1.7
MST ($\epsilon=1$)	0.23	$\times 1.0$	0.17	0.24	0.20	0.23	0.13	0.22	0.23
MST ($\epsilon=1000$)	0.18	$\times 1.3$	0.14	0.20	0.15	0.18	0.20	0.17	0.18

practitioner that is reassuring (risk is chosen at synthesis time, not imposed later by an adversary’s ingenuity), but it sharpens rather than settles the question of what to *do*. The metrics in widest use today each see only a slice of what a learning adversary recovers, and none of them see the *shape* of the risk: who is exposed, which attribute leaks, and whether the leakage is memorization a mechanism could remove or distributional inference no useful release can avoid. We close by translating these findings into auditing practice and marking what they leave open: in the spirit of an SoK, as much a map of what is not yet known as of what is.

5.1 Guidance for Privacy Auditing

Audit with a learning adversary, not a matching one. The standard of practice for disclosure control remains matching-based: equivalence-class risk in the El Emam tradition [22] and the similarity/distance scores shipped with synthetic-data tooling. Our calibrated ladder (Section 4.1) shows these understate exposure widely: the El Emam score rates CellSuppression at 0.42 where a one-line nearest-neighbor attack recovers 64%, because learning attacks generalize *across* equivalence classes rather than requiring exact QI matches; Ganey and De Cristofaro [25] reach the same verdict for similarity metrics even under differential privacy. The recommendation (already operationalized by the NIST CRC) is to audit with the strongest available attribute-inference attack, not a distributional-distance proxy. Frameworks are converging on attack-based evaluation [27, 33, 58]; our taxonomy and benchmark contribute a ranked attack suite, with CoBP-RA a strong, inexpensive default.

Report the distribution of risk, not its mean. A single mean R_{adv} conceals the two things a regulator most needs. The first is *mechanism*: our memorization test (motivated by *narcissus resiliency* [15]) separates individual memorization, which a better mechanism can remove, from distributional inference, which no informative release can. The second is *incidence*: high-fidelity synthesis can post a reassuring mean while exposing demographically rare individuals several-fold (Section 4.3), and because that concentration is a property of the release, it cannot be audited away with a weaker attacker. Both diagnostics (the train-versus-holdout gap and per-subgroup, per-feature scores) fall out of the same attack run at no extra cost; we recommend reporting both as standard.

Reconstruction and membership are complementary, and reconstruction is what the law asks about. Synthetic-data auditing has

standardized on membership inference [10, 57], but the two answer different questions and, as Section 4.2 shows, flag different individuals: roughly a third of records are scored discordantly. Membership asks whether a person was in the data; reconstruction asks what can be recovered about them, and it is the latter that HIPAA Expert Determination and the GDPR’s identifiability standard turn on: disclosure of attribute *values*, not of participation. A complete audit should run both, and treat reconstruction (not membership) as the measure aligned with the legal standard it certifies.

5.2 What Our Results Do and Do Not Say About Mechanisms

That de-identification carries no worst-case bound is not new [42, 62]; our contribution is to make its failure *measurable* and to characterize its shape. The leakage of CellSuppression and RankSwap is record-retention memorization (visible directly as a large train-holdout gap), and it is borne unevenly, falling on the majority and partially-retained minority groups while perfectly protecting the fully-suppressed. An auditor with our diagnostics can see this in a specific release rather than infer it from first principles.

Differential privacy is the only family in our study that is at once memorization-free and equitable across subgroups (the privacy mirror of Bagdasaryan et al. [6]’s disparate-*utility* finding, since the same noise that degrades a rare subgroup’s utility also denies an attacker its sharp conditional structure), and we take that, rather than any single reconstruction number, to be the property a deployment should weigh most heavily. We are deliberately careful about the privacy budget. Reconstruction risk does not climb with ϵ for MST the way membership risk has been shown to [28]; a release’s reconstruction exposure under MST appears governed as much by the model’s marginal structure as by its noise. But we swept ϵ only for MST and evaluated AIM at a single budget, so this is an observation about one mechanism, not a general law, and we resist reading a universal “ceiling” into it. What we can say stands on its own: among the generators we tested, only DP paired a formal guarantee with equitable, memorization-free releases in practice, whereas non-private deep generators (ARF, CTGAN, TVAE) avoided memorization only empirically and cannot promise it.

5.3 Limitations and Open Problems

Scope of the threat model. Our findings cover the two branches we study and should not be read as claims about the others: model

inversion, where the adversary holds the trained model, or reconstruction from aggregate statistics, the setting of the Census reconstruction theorem [18, 26].

An open taxonomy. Figure 2 organizes the attacks that *exist*: a snapshot, not a closed system. We populate several of its leaves ourselves (the message-passing CoBP-RA and the conditioned-generative CondMST, MultiHeadMLP, and partial-diffusion variants), but a taxonomy of known methods is an invitation rather than a census: new leaves will grow as existing ideas are refined, and genuinely new methodologies may earn branches we have not drawn. We offer it as scaffolding for that growth, not a final account.

The privacy budget across mechanisms. Our ϵ sweep covers MST; AIM and the other DP synthesizers were evaluated at a single budget. Whether reconstruction’s insensitivity to ϵ (in pointed contrast to membership inference) holds beyond MST is open, and is the cleanest next experiment our framework invites.

Continuous-feature synthesis. The continuous case (California, Section 3) is evaluated with a subset of attacks under NRMSE. It raises distinct challenges: DP discretization destroys utility at low ϵ , GANs and VAEs struggle with sharp continuous marginals, and there is no clean rarity-weighting analog for real-valued outputs. A complete treatment remains open.

Access, quasi-identifiers, and scale. Three axes we held fixed each bound our optimism. All our attacks are *black-box*; white-box access to model weights or training metadata (MST’s selected marginals, TabDDPM’s score function) would enable stronger attacks. We fix the *QI set*, yet Section 3.3 shows its composition strongly shapes risk, so a mis-specified audit QI can badly mis-estimate exposure: robust auditing under QI uncertainty is open. And our largest benchmark has 130 features, while genomics, EHR, and transaction data reach thousands; several of our strongest attacks (CoBP-RA, CondMST, diffusion) scale poorly in feature count, leaving efficient high-dimensional auditing an open systems problem.

From R_{adv} to legal thresholds. R_{adv} captures far more exposure than classical metrics (Section 4.1), but does not yet map onto HIPAA Expert Determination or GDPR identifiability. Formalizing that relationship is the bridge regulated deployments most need.

Toward a standard. The natural endpoint is a standardized, attack-based auditing benchmark: a fixed, ranked suite of reconstruction and membership attacks with agreed reporting (mean exposure, the memorization gap, and per-subgroup incidence) that a data holder at an agency such as NIST or the Census Bureau could run before release. Our taxonomy, attacks, and metrics are a step toward it; what remains is consensus on protocol, not a question of possibility.

Acknowledgments

This research was supported by NSF #2451163, NIH 1OT2OD032581, NSF NAIRR 240485 (Cloudbank AWS), and NSF NAIRR 240091 (TACC Frontera). Steven Golob is supported by an NSF CSGrad4US Fellowship. Sikha Pentiyala is affiliated with the eScience Institute.

We would like to thank the organizers of the NIST Privacy Collaborative Research Cycle, Christine Task and Gary Howarth, for designing and hosting the competition that motivated this work.

Ethical Considerations

This work develops and evaluates attacks, but its purpose is defensive: to measure reconstruction risk so that data holders and auditors can quantify and reduce it. All experiments use established public benchmark datasets (Adult, IPUMS census microdata, California Housing, NIST SBO, and CDC BRFSS); we collect no new personal data, and every “target” is a record in an already-public research dataset. The attacks are run only against synthetic and de-identified releases that we ourselves generated from these public datasets (never against a deployed system, a live data product, or any real individual), and the methodology was developed within the NIST Privacy Collaborative Research Cycle, a sanctioned red-team/blue-team program. We acknowledge the dual-use nature of attack research: the same techniques could in principle aid an adversary. We judge the disclosure benefit to outweigh this risk, because the attacks operate only on public benchmarks, we report no vulnerability in any fielded system, and every result is paired with auditing guidance and with the finding that differential privacy mitigates the exposure we measure.

Open Science

In keeping with PoPETs’ open-science policy, all code for the attacks, generators, experiment harness, and evaluation is publicly released, along with the configuration files and hyperparameters (Appendix B) needed to reproduce every table in this paper. All five datasets are publicly available from their original sources (Table 3). **Code:** <https://github.com/steveng9/Reconstruction>

AI Use

The authors used Claude Sonnet 4.6 (Anthropic) to revise manuscript text, install external libraries, run experiments, and query result databases. All AI-generated text was reviewed, edited for accuracy, and verified against experimental results by the authors. All scientific judgments, experimental designs, original NIST CRC participation and interpretations are the authors’ own. We have manually verified and are responsible for the accuracy, originality, and integrity of the output of all AI-based tools.

References

- [1] John M Abowd. 2018. The U.S. Census Bureau Adopts Differential Privacy. In *Proceedings of the 24th ACM SIGKDD International Conference on Knowledge Discovery & Data Mining*. 2867.
- [2] John M Abowd, Robert Ashmead, Ryan Cumings-Menon, Simson Garfinkel, Micah Heineck, Christine Heiss, Robert Johns, Daniel Kifer, Philip Leclerc, Ashwin Machanavajjhala, Brett Moran, William Sexton, Matthew Spence, and Pavel Zhuravlev. 2022. The 2020 Census Disclosure Avoidance System TopDown Algorithm. *Harvard Data Science Review* (2022). doi:10.1162/99608f92.529e3cb9 Special Issue 2.
- [3] Tristan Allard, Louis Béziaud, and Sébastien Gambs. 2023. SNAKE challenge: Sanitization algorithms under attack. In *Proceedings of the 32nd ACM international conference on information and knowledge management*. 5010–5014.
- [4] Meenatchi Sundaram Muthu Selva Annamalai, Andrea Gadotti, and Luc Rocher. 2024. A linear reconstruction approach for attribute inference attacks against synthetic data. In *33rd USENIX security symposium (USENIX security 24)*. 2351–2368.
- [5] Samuel A. Assefa, Danial Dervovic, Mahmoud Mahfouz, Robert E. Tillman, Prashant Reddy, and Manuela Veloso. 2020. Generating Synthetic Data in Finance:

- Opportunities, Challenges and Pitfalls. In *Proceedings of the First ACM International Conference on AI in Finance (ICAIF)*. ACM. doi:10.1145/3383455.3422554
- [6] Eugene Bagdasaryan, Omid Poursaeed, and Vitaly Shmatikov. 2019. Differential privacy has disparate impact on model accuracy. *Advances in neural information processing systems* 32 (2019).
 - [7] Barry Becker and Ronny Kohavi. 1996. Adult Data Set. <https://archive.ics.uci.edu/ml/datasets/adult>. doi:10.24432/C5XW20 UCI Machine Learning Repository.
 - [8] Gary Benedetto, Jordan C. Stanley, and Evan Totty. 2018. *The Creation and Use of the SIPP Synthetic Beta v7.0*. Working Paper. U.S. Census Bureau.
 - [9] Leo Breiman. 2001. Random Forests. *Machine Learning* 45, 1 (2001), 5–32.
 - [10] Nicholas Carlini, Steve Chien, Milad Nasr, Shuang Song, Andreas Terzis, and Florian Tramer. 2022. Membership inference attacks from first principles. In *2022 IEEE Symposium on Security and Privacy (SP)*. IEEE, 1897–1914.
 - [11] Nicholas Carlini, Jamie Hayes, Milad Nasr, Matthew Jagielski, Vikash Sehwal, Florian Tramer, Borja Balle, Daphne Ippolito, and Eric Wallace. 2023. Extracting Training Data from Diffusion Models. In *32nd USENIX Security Symposium (USENIX Security 23)*. 5253–5270.
 - [12] Nicholas Carlini, Florian Tramer, Eric Wallace, Matthew Jagielski, Ariel Herbert-Voss, Katherine Lee, Adam Roberts, Tom Brown, Dawn Song, Ulfar Erlingsson, Alina Oprea, and Colin Raffique. 2021. Extracting Training Data from Large Language Models. In *30th USENIX Security Symposium (USENIX Security 21)*. 2633–2650.
 - [13] Centers for Disease Control and Prevention. 2015. CDC Diabetes Health Indicators Dataset. <https://archive.ics.uci.edu/dataset/891/cdc+diabetes+health+indicators>. Behavioral Risk Factor Surveillance System (BRFSS).
 - [14] Chen Chen, Lingjuan Lyu, Han Yu, and Gang Chen. 2024. Practical Attribute Reconstruction Attack Against Federated Learning. *IEEE Transactions on Big Data* 10, 6 (2024), 851–863. doi:10.1109/TBDATA.2022.3159236
 - [15] Edith Cohen, Haim Kaplan, Yishay Mansour, Shay Moran, Kobbi Nissim, Uri Stemmer, and Eliad Tsfadia. 2024. Data reconstruction: When you see it and when you don't. *arXiv preprint arXiv:2405.15753* (2024).
 - [16] Hakan Demirtas. 2018. Flexible imputation of missing data. *Journal of statistical software* 85 (2018), 1–5.
 - [17] Travis Dick, Cynthia Dwork, Michael Kearns, Terrance Liu, Aaron Roth, Giuseppe Vietri, and Zhiwei Steven Wu. 2023. Confidence-ranked reconstruction of census microdata from published statistics. *Proceedings of the National Academy of Sciences* 120, 8 (2023), e2218605120.
 - [18] Irit Dinur and Kobbi Nissim. 2003. Revealing information while preserving privacy. In *Proceedings of the twenty-second ACM SIGMOD-SIGACT-SIGART symposium on Principles of database systems*. 202–210.
 - [19] Cynthia Dwork, Frank McSherry, Kobbi Nissim, and Adam Smith. 2006. Calibrating Noise to Sensitivity in Private Data Analysis. In *Proceedings of the Third Conference on Theory of Cryptography (TCC)*. Springer, 265–284.
 - [20] Cynthia Dwork and Aaron Roth. 2014. The algorithmic foundations of differential privacy. *Foundations and Trends® in Theoretical Computer Science* 9, 3–4 (2014), 211–407.
 - [21] Cynthia Dwork, Adam Smith, Thomas Steinke, and Jonathan Ullman. 2017. Exposed! a survey of attacks on private data. *Annual Review of Statistics and Its Application* 4, 1 (2017), 61–84.
 - [22] Khaled El Emam, Elizabeth Jonker, Luk Arbuckle, and Bradley Malin. 2011. A systematic review of re-identification attacks on health data. *PLoS One* 6, 12 (2011), e28071.
 - [23] Khaled El Emam, Lucy Mosquera, and Jason Bass. 2020. Evaluating Identity Disclosure Risk in Fully Synthetic Health Data: Model Development and Validation. *Journal of Medical Internet Research* 22, 11 (2020), e23139. doi:10.2196/23139
 - [24] Matt Fredrikson, Somesh Jha, and Thomas Ristenpart. 2015. Model Inversion Attacks that Exploit Confidence Information and Basic Countermeasures. In *Proceedings of the 22nd ACM SIGSAC Conference on Computer and Communications Security*. 1322–1333.
 - [25] Georgi Ganey and Emiliano De Cristofaro. 2025. The Inadequacy of Similarity-Based Privacy Metrics: Privacy Attacks Against “Truly Anonymous” Synthetic Datasets. In *2025 IEEE Symposium on Security and Privacy (SP)*. 4007–4025. doi:10.1109/SP61157.2025.00218 ISSN: 2375-1207.
 - [26] Simson L Garfinkel, John M Abowd, and Christian Martindale. 2019. Understanding Database Reconstruction Attacks on Public Data. *Commun. ACM* 62, 3 (2019), 46–53.
 - [27] Matteo Gioni, Franziska Boenisch, Christoph Wehmeyer, and Borbála Tasnádi. 2023. A Unified Framework for Quantifying Privacy Risk in Synthetic Data. *Proceedings on Privacy Enhancing Technologies* (2023).
 - [28] Steven Golob, Sikha Pentiyala, Anuar Maratkhan, and Martine De Cock. 2025. Privacy Vulnerabilities in Marginals-based Synthetic Data. In *2025 IEEE Conference on Secure and Trustworthy Machine Learning (SaTML)*. 977–995. doi:10.1109/SaTML64287.2025.00059
 - [29] Jamie Hayes, Luca Melis, George Danezis, and Emiliano De Cristofaro. 2019. LOGAN: Membership inference attacks against generative models. *Proceedings on Privacy Enhancing Technologies* 1 (2019), 133–152.
 - [30] Benjamin Hilprecht, Martin Härterich, and Daniel Bernau. 2019. Monte carlo and reconstruction membership inference attacks against generative models. *Proceedings on Privacy Enhancing Technologies* 2019, 4 (2019), 232–249.
 - [31] Jonathan Ho, Ajay Jain, and Pieter Abbeel. 2020. Denoising Diffusion Probabilistic Models. In *Advances in Neural Information Processing Systems*, Vol. 33. 6840–6851.
 - [32] Noah Hollmann, Samuel Müller, Katharina Eggenberger, and Frank Hutter. 2023. TabPFN: A Transformer That Solves Small Tabular Classification Problems in a Second. In *International Conference on Learning Representations (ICLR)*.
 - [33] Florimond Houssiau, James Jordon, Samuel N Cohen, Owen Daniel, Andrew Elliott, James Geddes, Callum Mole, Camila Rangel-Smith, and Lukasz Szpruch. 2022. TAPAS: a toolbox for adversarial privacy auditing of synthetic data. *arXiv preprint arXiv:2211.06550* (2022).
 - [34] Yuzheng Hu, Fan Wu, Qibin Li, Yunhui Long, Gonzalo Garrido, Chang Ge, Bolin Ding, David Forsyth, Bo Li, and Dawn Song. 2024. SoK: privacy-preserving data synthesis. In *2024 IEEE Symposium on Security and Privacy (SP)*. 4696–4713.
 - [35] Bargav Jayaraman and David Evans. 2022. Are attribute inference attacks just imputation?. In *Proceedings of the 2022 ACM SIGSAC conference on computer and communications security*. 1569–1582.
 - [36] James Jordon, Lukasz Szpruch, Florimond Houssiau, Mirko Bottarelli, Giovanni Cherubin, Carsten Maple, Samuel N Cohen, and Adrian Weller. 2022. Synthetic Data—what, why and how? *arXiv preprint arXiv:2205.03257* (2022).
 - [37] Guolin Ke, Qi Meng, Thomas Finley, Taifeng Wang, Wei Chen, Weidong Ma, Qiwei Ye, and Tie-Yan Liu. 2017. Lightgbm: A highly efficient gradient boosting decision tree. *Advances in neural information processing systems* 30 (2017).
 - [38] Akim Kotelnikov, Dmitry Baranchuk, Ivan Rubachev, and Artem Babenko. 2023. Tabddpm: Modelling tabular data with diffusion models. In *International conference on machine learning*. PMLR, 17564–17579.
 - [39] Zinan Lin, Ashish Khetan, Giulia Fanti, and Sewoong Oh. 2018. PacGAN: The power of two samples in generative adversarial networks. In *Advances in Neural Information Processing Systems*, Vol. 31.
 - [40] Andreas Lugmayr, Martin Danelljan, Andres Romero, Fisher Yu, Radu Timofte, and Luc Van Gool. 2022. Repaint: Inpainting using denoising diffusion probabilistic models. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*. 11461–11471.
 - [41] Lingjuan Lyu and Chen Chen. 2021. A Novel Attribute Reconstruction Attack in Federated Learning. *arXiv preprint arXiv:2108.06910* (2021).
 - [42] Ashwin Machanavajjhala, Daniel Kifer, Johannes Gehrke, and Muthuramakrishnan Venkatasubramanian. 2007. *t*-Diversity: Privacy Beyond *k*-Anonymity. *ACM Transactions on Knowledge Discovery from Data* 1, 1 (2007), 3.
 - [43] Ryan McKenna, Gerome Miklau, and Daniel Sheldon. 2021. Winning the NIST Contest: A scalable and general approach to differentially private synthetic data. *arXiv preprint arXiv:2108.04978* (2021).
 - [44] Ryan McKenna, Brett Mullins, Daniel Sheldon, and Gerome Miklau. 2022. Aim: An adaptive and iterative mechanism for differentially private synthetic data. *arXiv preprint arXiv:2201.12677* (2022).
 - [45] Ryan McKenna, Daniel Sheldon, and Gerome Miklau. 2019. Graphical-model based estimation and inference for differential privacy. In *International conference on machine learning*. PMLR, 4435–4444.
 - [46] Matthieu Meeus, Florent Guepin, Ana-Maria Crețu, and Yves-Alexandre de Montjoye. 2023. Achilles’ heels: vulnerable record identification in synthetic data publishing. In *European symposium on research in computer security*. Springer, 380–399.
 - [47] Richard A. Moore. 1996. *Controlled data-swapping techniques for masking public use microdata sets*. Technical Report RR 96/04. U.S. Bureau of the Census, Statistical Research Division.
 - [48] Krishnamurthy Muralidhar and Josep Domingo-Ferrer. 2023. Database reconstruction is not so easy and is different from reidentification. *Journal of Official Statistics* 39, 3 (2023), 381–398. Publisher: SAGE Publications Sage UK: London, England.
 - [49] Arvind Narayanan and Vitaly Shmatikov. 2008. Robust De-Anonymization of Large Sparse Datasets. In *2008 IEEE Symposium on Security and Privacy (SP)*. IEEE, 111–125.
 - [50] National Institute of Standards and Technology. 2018. NIST Differential Privacy Synthetic Data Challenge (Arizona Dataset). <https://www.nist.gov/ct/pscr/open-innovation-prize-challenges>. Synthetic census-like dataset.
 - [51] National Institute of Standards and Technology. 2018. NIST Synthetic Data Challenge: Survey of Business Owners (SBO). <https://www.nist.gov/ct/pscr/open-innovation-prize-challenges>. Synthetic business transaction dataset.
 - [52] Beata Nowok, Gillian M Raab, and Chris Dikken. 2016. synthpop: Bespoke creation of synthetic data in R. *Journal of statistical software* 74 (2016), 1–26.
 - [53] Jesse Read, Bernhard Pfahringer, Geoff Holmes, and Eibe Frank. 2011. Classifier chains for multi-label classification. *Machine Learning* 85, 3 (2011), 333–359. doi:10.1007/s10994-011-5256-5
 - [54] Luc Rocher, Julien M Hendrickx, and Yves-Alexandre De Montjoye. 2019. Estimating the success of re-identifications in incomplete datasets using generative models. *Nature Communications* 10, 1 (2019), 3069.
 - [55] Pierangela Samarati. 2001. Protecting Respondents’ Identities in Microdata Release. *IEEE Transactions on Knowledge and Data Engineering* 13, 6 (2001), 1010–1027. doi:10.1109/69.971193

- [56] scikit-learn developers. 2024. California Housing Dataset. https://scikit-learn.org/stable/modules/generated/sklearn.datasets.fetch_california_housing.html. Derived from 1990 US Census data.
- [57] Reza Shokri, Marco Stronati, Congzheng Song, and Vitaly Shmatikov. 2017. Membership inference attacks against machine learning models. In *2017 IEEE Symposium on Security and Privacy (SP)*. IEEE, 3–18.
- [58] Theresa Stadler, Bristena Oprisanu, and Carmela Troncoso. 2022. Synthetic data—anonimisation groundhog day. In *31st USENIX security symposium (USENIX security 22)*. 1451–1468.
- [59] Daniel J Stekhoven and Peter Bühlmann. 2012. MissForest—non-parametric missing value imputation for mixed-type data. *Bioinformatics* 28, 1 (2012), 112–118.
- [60] Jonathan AC Sterne, Ian R White, John B Carlin, Michael Spratt, Patrick Royston, Michael G Kenward, Angela M Wood, and James R Carpenter. 2009. Multiple imputation for missing data in epidemiological and clinical research: potential and pitfalls. *Bmj* 338 (2009).
- [61] Latanya Sweeney. 2000. *Simple Demographics Often Identify People Uniquely*. Technical Report LIDAP-WP4. Carnegie Mellon University, Data Privacy Laboratory.
- [62] Latanya Sweeney. 2002. k-anonymity: A model for protecting privacy. *International Journal of Uncertainty, Fuzziness and Knowledge-Based Systems* (2002).
- [63] Matthias Templ, Alexander Kowarik, and Bernhard Meindl. 2015. Statistical Disclosure Control for Micro-Data Using the R Package sdcMicro. *Journal of Statistical Software* 67, 4 (2015), 1–36.
- [64] U.S. Department of Health and Human Services, Office for Civil Rights. 2012. Guidance Regarding Methods for De-identification of Protected Health Information in Accordance with the HIPAA Privacy Rule. <https://www.hhs.gov/hipaa/for-professionals/special-topics/de-identification/index.html>.
- [65] Boris Van Breugel, Hao Sun, Zhaozhi Qian, and Mihaela van der Schaar. 2023. Membership inference attacks against synthetic data through overfitting detection. *arXiv preprint arXiv:2302.12580* (2023).
- [66] Stef Van Buuren and Karin Groothuis-Oudshoorn. 2011. mice: Multivariate imputation by chained equations in R. *Journal of statistical software* 45 (2011), 1–67.
- [67] Jason Walonoski, Mark Kramer, Joseph Nichols, Andre Quina, Chris Moesel, Dylan Hall, Carlton Duffett, Kudakwashe Dube, Thomas Gallagher, and Scott McLachlan. 2018. Synthea: An approach, method, and software mechanism for generating synthetic patients and the synthetic electronic health care record. *Journal of the American Medical Informatics Association* 25, 3 (2018), 230–238. doi:10.1093/jamia/ocx079
- [68] David S Watson, Kristin Blesch, Jan Kapar, and Marvin N Wright. 2023. Adversarial random forests for density estimation and generative modeling. In *International conference on artificial intelligence and statistics*. PMLR, 5357–5375.
- [69] Leon Willenborg and Ton de Waal. 2001. *Elements of Statistical Disclosure Control*. Springer, New York.
- [70] Lei Xu, Maria Skoularidou, Alfredo Cuesta-Infante, and Kalyan Veeramachaneni. 2019. Modeling tabular data using conditional gan. *Advances in neural information processing systems* 32 (2019).
- [71] M Yaghini, B Kulynych, G Cherubin, M Veale, and C Troncoso. 2022. Disparate vulnerability to membership inference attacks. *Proceedings on Privacy Enhancing Technologies* 2022, 1 (2022), 460–480.
- [72] Samuel Yeom, Irene Giacomelli, Matt Fredrikson, and Somesh Jha. 2018. Privacy Risk in Machine Learning: Analyzing the Connection to Overfitting. In *2018 IEEE 31st Computer Security Foundations Symposium (CSF)*. IEEE, 268–282.
- [73] Jinsung Yoon, James Jordon, and Mihaela Schaar. 2018. Gain: Missing data imputation using generative adversarial nets. In *International conference on machine learning*. PMLR, 5689–5698.
- [74] Jun Zhang, Graham Cormode, Cecilia M Procopiuc, Divesh Srivastava, and Xiaokui Xiao. 2017. PrivBayes: Private data release via Bayesian networks. *ACM Transactions on Database Systems (TODS)* 42, 4 (2017), 1–41.
- [75] Benjamin Zi Hao Zhao, Aviral Agrawal, Catisha Coburn, Hassan Jameel Asghar, Raghav Bhaskar, Mohamed Ali Kaafar, Darren Webb, and Peter Dickinson. 2021. On the (in) feasibility of attribute inference attacks on machine learning models. In *2021 IEEE european symposium on security and privacy (EuroS&P)*. IEEE, 232–251.

A Preliminaries

A.1 Dataset Details

The five datasets (summarized in Table 3) span the data domains and structural regimes most relevant to disclosure control; all are public benchmarks. **Adult** [7] is the 1994 US Census income extract, a standard machine-learning, fairness, and privacy benchmark of mixed demographic and employment features. **NIST Arizona** [50]

is IPUMS 1940 Arizona census microdata; we use the 25-feature subset released for the NIST CRC, with the competition’s two quasi-identifier sets (QI_1, QI_2 ; Table 25). **California Housing** [56] is 1990 census block-group housing data and our only fully-continuous dataset, scored by NRMSE rather than rarity-weighted accuracy. **NIST SBO** [51] is Survey-of-Business-Owners public-use microdata, our most feature-rich dataset (130 mostly-binary firm-level financial and demographic attributes). **CDC Diabetes** [13] is the CDC BRFSS Diabetes Health Indicators set, 22 mostly-binary health features. Per-dataset quasi-identifier memberships are tabulated in Appendix E.8 (Tables 23–27).

A.2 De-identification and Synthetic Data Generation Methods

The nine methods studied span three broad families. *Statistical disclosure limitation (SDL)* methods transform or suppress real records without generating new ones. *Statistical synthesis* methods fit explicit probabilistic models to the data and draw new records from those models; the DP variants add calibrated noise to the fitted statistics. *Deep generative synthesis* methods train neural networks to learn the data distribution and sample from the learned model. Only MST and AIM carry a formal differential-privacy guarantee; all others are non-DP. The exact generation settings for every method (library defaults except where noted) are consolidated in Appendix B.

Statistical Disclosure Limitation (non-synthetic, non-DP)

RankSwap. Records are sorted by a continuous key feature, partitioned into consecutive bins of fixed size, and within each bin the values of designated sensitive numerical columns are permuted among records [47, 63]. All non-swapped columns are released verbatim, so the output consists of *real* records with certain attributes replaced. RankSwap preserves marginal distributions of the swapped features but breaks individual-level linkages. In our experiments the swapped columns are the continuous / high-cardinality numeric attributes for each dataset (e.g., for **ADULT**: age, fnlwgt, education-num, capital-gain, capital-loss, hours-per-week).

CellSuppression. Records whose quasi-identifier combination falls into a group with fewer than k members are removed entirely; the remaining records are released unmodified [69]. This enforces a form of k -anonymity on the released microdata. The suppressed fraction can be substantial for datasets with many rare strata. In our experiments we use $k = 5$ and key variables that form the quasi-identifier of interest for each dataset (e.g., for **ADULT**: age, workclass, education, sex, race, native-country). Because suppression reduces record count, the released file is smaller than the original training set; attacks on CellSuppression synth operate on a reduced candidate pool.

Statistical Synthesis — Differentially Private

MST. Maximum Spanning Tree [43], the winning entry in the 2018 NIST Differential Privacy Synthetic Data Challenge. Selects a set of pairwise marginals by computing a maximum spanning tree of mutual information scores, measures those marginals with calibrated Gaussian noise to achieve (ϵ, δ) -DP, then fits a graphical

model (via Private-PGM [45]) and samples synthetic records from it. We sweep $\epsilon \in \{0.1, 1, 10, 100, 1000\}$.

AIM.. Adaptive and Iterative Mechanism [44]. Extends MST with an online marginal-selection strategy: rather than fixing the marginal set upfront, AIM iteratively measures the marginal that reduces model error the most, allocating the privacy budget adaptively. The same Private-PGM backend is used for graphical-model fitting and sampling. AIM achieves higher utility than MST at matched ϵ because it targets the most informative statistics first. We use $\epsilon \in \{1, 3, 10\}$.

Statistical Synthesis — Non-DP

Synthpop. A sequential synthesis method implemented in the synthpop R package [52]. Each feature is synthesised in turn by fitting a conditional model (default: CART) given all previously synthesised columns. Captures complex non-linear conditional distributions. No differential-privacy guarantee; widely used in official statistics and social-science disclosure control.

Deep Generative Synthesis — Non-DP

CTGAN.. Conditional Tabular GAN [70]. Adversarially trained: a conditional generator (conditioned on a randomly sampled categorical column to cover the full marginal) competes against a PacGAN discriminator [39] that operates on pairs of records to reduce mode collapse. Continuous columns are transformed via mode-specific normalisation. Training can be unstable; CTGAN tends to excel on datasets with many categorical columns and struggle on continuous-heavy ones.

TVAE.. Tabular Variational Autoencoder [70]. An encoder maps records to a latent Gaussian, and a decoder reconstructs synthetic records by sampling and decoding. Uses the same mode-specific normalisation as CTGAN but is not adversarially trained, yielding more stable optimisation at the cost of potentially under-representing rare modes.

ARF.. Adversarial Random Forests [68]. Iteratively trains a random forest to classify real versus synthetic records; the forest’s leaf partitions define a density estimate from which new synthetic data are drawn. A hybrid between ensemble methods and generative modelling — no neural networks or minimax game. The “adversarial” label refers to the use of a classifier to drive density estimation, not a GAN-style training objective.

TabDDPM.. Denoising Diffusion Probabilistic Model for tabular data [38]. Learns to reverse a Gaussian noise process applied to continuous feature embeddings (multinomial diffusion for categorical features). At inference, starts from noise and iteratively denoises to produce synthetic records. Achieves state-of-the-art fidelity on many tabular benchmarks and, in our experiments, tends to produce the highest-vulnerability synthetic data.

A.3 Membership Inference

Membership inference (MIA) asks whether a candidate record x was in the training set D_{train} used to fit the generator. An attack assigns x a real-valued *membership score* $s(x)$, higher meaning more member-like, and a decision is obtained by thresholding s . We evaluate

three scores against the synthetic release D_{syn} . *SynthDistance* [29] sets $s(x) = -\min_{y \in D_{\text{syn}}} d(x, y)$: members tend to lie closer to the synthetic data the generator produced. *NNDR* (nearest-neighbour distance ratio) [58] normalizes that distance by the distance to the nearest leave-one-out *training* neighbour, which down-weights records that are close to everything. *RA-as-MIA* (Section 4.2) instead uses reconstruction quality itself as the score: how well the full reconstruction attack recovers x ’s hidden features from its QI.

Because every score is a ranking, we report threshold-free metrics: ROC AUC (Table 8); the *membership advantage* [72], TPR—FPR at the best threshold; the **TPR at low FPR** (FPR = 0.01), which captures confident re-identification of a few individuals [10]; and balanced accuracy. Members and non-members are drawn from disjoint training samples (Section 4.2).

B Hyperparameters

Unless noted, every attack and generator runs at the fixed defaults below; the experiment scripts log the full effective parameter set to WandB for each run.

Attacks. The per-feature classifiers use scikit-learn / LightGBM defaults except: KNN uses $k=1$ with distance weighting; the random forest uses 25 trees at maximum depth 25; LightGBM uses 100 boosting rounds; logistic regression 100 iterations; the SVM an RBF kernel ($C=1$). The MLP is a single 300-unit hidden layer (dropout 0.2, batch 264, Adam at 3×10^{-4} , up to 500 epochs with patience 60); MultiHeadMLP shares one $1000 \rightarrow 500$ trunk with a softmax head per feature; ARFFormer uses 2 layers, 4 heads, and embedding width 64. CoBP-RA wraps the same 25-tree forest with belief propagation over an MST of pairwise synthetic marginals (local PMI from the 100 nearest synthetic neighbours, Laplace smoothing $\alpha=10^{-6}$); its continuous variant quantile-bins each hidden feature into 20 bins and switches to global PMI. The diffusion attacks (CondDDPM, CondRePaint, CondDDPMWithMLP) share one MLP backbone of layer widths [512, 1024, 1024, 1024, 1024, 512] trained for 200,000 steps over 2,000 diffusion timesteps (batch 4096, lr 6×10^{-4}); CondMST fits a graphical model with 20 ordinal bins per continuous column, with each hidden feature’s clique bounded to size $k=3$ (joined to its two highest mutual-information QI columns) and sampled in default sample mode. LinearReconstruction uses 3-way marginal queries solved with Gurobi.

Generators. All nine SDG methods run at their library defaults, with three documented exceptions. MST and AIM are swept over $\epsilon \in \{0.1, 1, 10, 100, 1000\}$ and $\{1, 3, 10\}$ respectively, with continuous columns pre-binned into ordinal bins so that the full privacy budget is spent on synthesis rather than on private bound estimation. RankSwap permutes the continuous / high-cardinality columns within rank blocks (Appendix A.2 lists the swapped columns per dataset), and CellSuppression enforces $k=5$ -anonymity on each dataset’s QI key variables. For the 130-column NIST SBO data the TabDDPM backbone is widened to [1024, 2048, 2048, 2048, 1024] and trained for 300,000 steps.

Generator implementations. For reproducibility we note the underlying libraries: MST and AIM use SmartNoise-Synth (snsynth); TVAE and CTGAN use SDV; ARF uses Synthcity’s arf plugin; TabDDPM uses the official tabular-diffusion implementation [38]. The

Table 10: Reconstruction accuracy (R_{adv} , %) on NIST SBO. Mode baseline = 32.1% throughout; bold = best per column among ML attacks. † = novel attack. (1k rows, QI_1 .)

Attack	MST ($\epsilon=0.1$)	MST ($\epsilon=1$)	MST ($\epsilon=10$)	MST ($\epsilon=100$)	MST ($\epsilon=1000$)	Cell Supp.	RankSwap	Synthpop	TVAE	CTGAN	ARF	TabDDPM
Mode	32.1	32.1	32.1	32.1	32.1	32.1	32.1	32.1	32.0	32.1	32.1	32.1
KNN	32.5	32.0	32.4	32.6	32.7	39.9	72.2	33.9	33.0	33.1	33.3	32.1
Naive Bayes	32.1	32.3	32.4	32.6	32.4	36.4	39.6	33.3	33.2	32.0	34.3	32.1
Random Forest	32.4	32.0	32.5	32.6	32.7	35.5	66.0	34.1	33.0	32.6	33.0	32.1
CoBP-RA [†]	32.1	32.0	32.9	33.5	33.3	42.3	80.8	34.6	33.1	34.3	33.5	32.1

de-identification and statistical-synthesis baselines call R through rpy2: Synthpop uses the synthpop package (CART engine), and RankSwap and CellSuppression use sdcMicro (rankSwap and localSuppression, respectively).

C Implementation Details

Three attacks turn an SDG method into a reconstruction primitive by conditioning its generative process on the target’s quasi-identifiers: fit (or reuse) a generator on the synthetic release, then sample the hidden features conditioned on each target’s known QI values. We describe the three adaptations below; the remaining attacks are the standard supervised models of Section 2.2.

C.1 Conditional MST (CondMST)

CondMST fits an MST graphical model (Private-PGM [45]) directly on the *synthetic* release rather than on private data, then samples each target’s hidden features conditioned on its observed QI . Fitting is *noiseless*: because the synthetic data is already public, we drive the mechanism’s Gaussian noise to zero ($\sigma \rightarrow 0$) so the model captures the release’s own pairwise-marginal structure as faithfully as possible. Continuous columns are discretized into 20 ordinal bins before fitting and decoded back to bin representatives at sampling time. For each target we clamp the QI factors to the observed values and draw the hidden features from the resulting conditional distribution (default sample mode; deterministic argmax and top- p truncation are also available). Several clique structures are possible: the *independent* variant fits a separate single-feature model per hidden feature, while the *bounded* and *hub* variants enlarge the cliques tying hidden features to their most informative QI columns. We report the *bounded* variant with maximum clique size $k=3$ (each hidden feature joined to its two highest mutual-information QI columns) as **CondMST** in Table 1, the strongest configuration we found.

C.2 Conditional TabDDPM (CondDDPM)

CondDDPM trains a TabDDPM diffusion model on the synthetic release with the QI columns supplied as conditioning, then generates the hidden columns for each target by running reverse diffusion conditioned on that target’s QI . CondRePaint reuses the identical trained model but replaces ancestral sampling with RePaint inpainting [40]: the known QI cells are held fixed (re-noised to the current

timestep) while the hidden cells are denoised, with repeated re-sampling jumps to harmonize the two. Because both attacks share one checkpoint, whichever runs second reuses the cached model. CondDDPMWithMLP adds a second stage: a per-feature MLP that maps the diffusion model’s hidden-feature “hints” (with the encoded QI) to a final prediction, correcting systematic biases in the raw diffusion output.

C.3 Belief-Propagation Correction (CoBP-RA)

CoBP-RA starts from independent per-feature posteriors and corrects them with the dependency structure the release exposes:

- (1) **Unary posteriors.** For each hidden feature fit a random forest on the synthetic data ($QI \rightarrow \text{feature}$) and read off the target’s posterior $P(h \mid x^Q)$.
- (2) **Dependency graph.** Build the maximum spanning tree of pairwise mutual information between hidden features, measured on the synthetic data.
- (3) **Pairwise potentials.** Weight each edge by the pointwise mutual information $PMI(h_a, h_b)$ of the synthetic joint table, estimated either globally or from the target’s 100 nearest synthetic neighbours.
- (4) **Message passing.** Run sum-product belief propagation (exact on the tree) to reconcile the unary posteriors with the pairwise couplings.
- (5) **Decode.** Predict each hidden feature by its posterior mode.

When the synthetic features are independent (all $PMI \approx 0$) the pairwise messages vanish and CoBP-RA reduces exactly to the underlying random forest; its gains come entirely from joint structure the release happens to preserve. The continuous variant quantile-bins each hidden feature, runs the same message passing on bin labels, and decodes to bin midpoints (Appendix E.9).

D Attack Enhancements: Ensembling and Chaining

Two natural strategies for strengthening reconstruction attacks present themselves when multiple hidden features must be predicted: ensembling the predictions of several attacks, and chaining predictions auto-regressively. Both played a role in our NIST CRC winning submission. In extensive controlled experiments against fixed SDG configurations (spanning more than 200 distinct (sample,

SDG, QI) combinations), however, neither enhancement reliably improves upon the best individual attack. We report this as an informative negative finding. The oracle analysis below reveals that the information required to improve does *exist* in the data; it simply cannot be accessed by any of the schemes we tried. This locates the problem precisely: the bottleneck is not attack sophistication but a structural property of reconstruction on synthetic data. Synthetic releases preserve population-level distributional structure; they are blind to which specific records are hardest to reconstruct. No voting or prediction scheme can overcome the fundamental ceiling set by how much higher-order joint structure the SDG retains beyond pairwise marginals. This theme runs through both ensembling and chaining, and we develop it in each subsection below.

D.1 Ensembling

The *Best ensemble* row of Table 1 is a single fixed configuration applied to every generator: **hard (majority) voting** over five base attacks: CoBP-RA (the primary), RandomForest, LightGBM, MLP, and KNN, each at its default hyperparameters (Appendix B). We selected it as the best-averaging combination among the configurations we swept; it nonetheless lands below CoBP-RA alone, as the oracle analysis below explains.

We exhaustively evaluated all $\binom{9}{2} = 36$ two-attack pairwise soft-voting ensembles, covering the nine leading attacks from Table 1 across four SDG methods (Synthpop, TabDDPM, TVAE, MST $\epsilon=10$). Figure 3 visualizes the pairwise grids for four panels (the average over all nine generators plus TabDDPM, MST $\epsilon=10$, and TVAE) with the diagonal giving each attack’s individual score and the off-diagonal its pairwise ensemble. For each of the 5 training samples and 4 SDG methods, we record the best pairwise ensemble RA against the best individual RA for that (sample, SDG) pair; across these 20 evaluation jobs, the mean gain was -1.2 RA points and only 5 of the 20 jobs showed any positive gain at all, with the maximum positive improvement being $+0.12$ pp. We also evaluated a meta-learning ensemble (stacking): a per-feature logistic regression trained on held-out synthetic-data predictions (out-of-fold cross-validation on the synthetic dataset, where some synthetic records are withheld when training each base model, and then those withheld predictions are used to train the meta-learner) augmented with QI features. Stacking outperformed soft voting but still fell below the best individual attack. The reason is distribution shift: the meta-model learns reliability weights (which base model to trust in which QI region) from synthetic data, but real training targets occupy different regions of feature space than synthetic records, so the learned confidence estimates do not transfer.

A third approach, *oracle-tuned weighted averaging*, grid-searched all 816 convex weight combinations (step size 0.1) over five attacks (CoBP-RA, LightGBM, MLP, KNN, NaiveBayes) across five samples and five SDG methods. The optimal weights on this grid were CoBP-RA=0.20, MLP=0.50, KNN=0.20, NaiveBayes=0.10, LightGBM=0.00, achieving 26.7% RA averaged over all SDG methods. This is identical to CoBP-RA alone (26.7%). Even with weights tuned directly on the test distribution (an oracle advantage no real attacker has), the optimal weighted combination provides no improvement over the single best attack. This closes the question of whether cleverer weighting schemes could rescue ensembling: the answer is no.

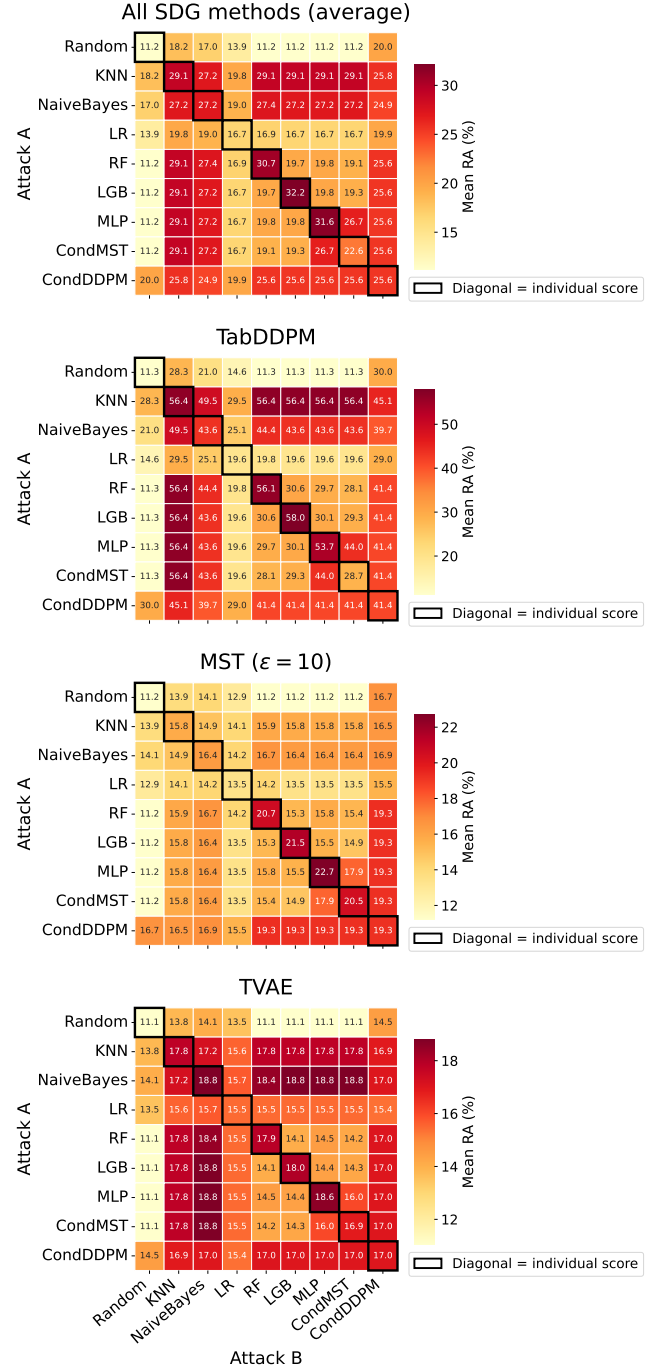


Figure 3: Pairwise ensemble reconstruction accuracy (R_{adv} mean), one panel per generator: the average over all nine SDG methods, TabDDPM, MST $\epsilon=10$, and TVAE. In each panel rows and columns index the individual attacks: diagonal cells (black border) give each attack’s individual score, off-diagonal cells the soft-voting ensemble of the corresponding pair; colour encodes R_{adv} (higher = better reconstruction). No off-diagonal pair improves meaningfully on the better of its two attacks. (Adult, 1k rows, QI_{demo}.)

The oracle analysis reveals why: if one could *always identify* which attack happened to predict correctly on a given record and always use that attack, the ensemble would achieve +11.4 RA points over the best individual. This gap confirms genuine complementarity; the attacks do not fail identically on every record. Yet the feature-level oracle (choosing the best attack *per feature type* rather than per individual record) adds only +2.7 pp. Together these two oracle bounds sharply locate the bottleneck: complementarity is record-level, not feature-level. Every attack struggles on the same hard records regardless of which feature is being predicted. No model combination can identify, from available QI signals, which records are in that hard-to-reconstruct region. This is a structural property: the synthetic data simply does not contain enough information to distinguish easy from hard individual records beyond the population-level signal the QI already provides. The pairwise-ensemble heatmaps (Figure 3) make this visible: no off-diagonal pair improves meaningfully on the better of its two constituents.

D.2 Chaining

The *Best chain* row of Table 1 is also a single fixed configuration: **CoBP-RA** as the base attack with **hard chaining** under a **mutual-information ordering** (hidden features predicted in decreasing mutual information with the QI), each predicted value fed back as an additional known feature. It was the strongest of the orderings and feedback modes we swept, and like the best ensemble it does not surpass CoBP-RA alone.

Chaining [53] predicts hidden features sequentially, feeding each predicted value back as an additional input for the next step. Formally, for hidden features ordered as $H = \{h_1, \dots, h_m\}$, the n -th predictor is trained as

$$\mathcal{A}_n = \mathcal{A}(D_{\text{syn}}[Q \cup \{h_1, \dots, h_{n-1}\}], D_{\text{syn}}[h_n]),$$

and at inference, predicted values are substituted for earlier features:

$$\hat{h}_n = \mathcal{A}_n(D_{\text{train}}[Q \cup \{\hat{h}_1, \dots, \hat{h}_{n-1}\}]).$$

We evaluate chain orderings by decreasing mutual information with Q , decreasing correlation, and randomly. Hard chaining (feeding back the most-likely predicted label) was essentially neutral across all configurations: the best result was +0.3 pp (RF, mutual-information ordering on Adult) and the MLP degraded by 1.9 pp. No ordering strategy produced consistent improvements.

Three additional mechanisms designed to reduce error propagation also failed: *soft chaining* (passing full posterior vectors instead of hard labels) degraded RF by 25.9 pp as noisy probabilities corrupted downstream conditioning; *confidence-gated chaining* (replacing low-confidence predictions with a uniform prior) recovered no performance at any threshold; and *Gibbs chaining* (iterating the chain over multiple passes) stabilized below non-chained RF for RF (locking in incorrect early assignments) and progressively degraded for MLP.

The oracle bound provides the definitive perspective: conditioning on *true* hidden-feature values at each step yields +7.2 pp for RF and only +1.5 pp for MLP. The MLP oracle gain being near zero is particularly striking: it means that even *perfect* previous predictions provide almost no additional signal beyond QI for the MLP’s remaining features. This rules out error propagation as the primary limitation. The synthetic joint distribution simply does

not preserve enough higher-order conditional structure (beyond what QI captures) for chaining to exploit. Notably, this holds for both MST (which captures only pairwise marginals by design; best chaining result on MST $\varepsilon=1000$: +0.1 pp over non-chained) and TabDDPM (which models the full joint via diffusion; best chaining result on TabDDPM: +0.2 pp). The near-zero gains on TabDDPM are especially instructive: despite learning a high-fidelity joint generative model, the synthetic data TabDDPM produces does not encode enough higher-order *conditional* structure (hidden|QI+previous-hidden) to make auto-regressive conditioning useful beyond the direct QI signal.

E Additional Results

E.1 NIST SBO Reconstruction Results

Table 10 reports reconstruction on NIST SBO at 1k rows under QI_1 (the per-feature structure behind the aggregate is discussed in Section 3). De-identification is again the exposed regime: rank swapping leaks heavily (CoBP-RA 80.8%, KNN 72.2%), because small firms form near-singleton QI groups that pass through the swap unchanged, and cell suppression follows (42.3%). Every generative and DP release, by contrast, sits at or within a point of the 32.1% Mode baseline (TabDDPM matches Mode exactly for all attacks, and the full MST ε sweep barely moves), so this 130-feature financial data leaves little structure reconstructable beyond the marginal. CoBP-RA is best or tied in nearly every column, consistent with the main benchmark.

E.2 Extended Linear Reconstruction Results

For the 5-category race target (Adult, 1k rows), the LP achieves 28.0% RA on average across SDG methods, essentially matching RF’s 27.1%; gains concentrate against TabDDPM (LP: 65.5%, RF: 64.2%). At 10k rows, LP scores 34.5% vs. RF’s 33.9%, a marginal difference that does not justify the substantially higher compute cost. For joint binary prediction (Adult 10k, jointly predicting income and sex as a $2^2=4$ -class problem), LP scores 72.1 vs. RF 74.0, with LP not winning against any individual SDG method. On CDC Diabetes (jointly predicting Diabetes_binary, Stroke, HighBP), RF beats LP at both 1k (59.4 vs. 58.1) and 100k (59.1 vs. 57.8) rows. In no setting does the LP surpass simple ML baselines.

E.3 Feature-Level Reconstruction Accuracy by MST ε

Table 13 reports per-feature R_{adv} averaged over five attacks (Naive-Bayes, KNN, RandomForest, CoBP-RA, MLP) across MST $\varepsilon \in \{0.1, 1, 10, 100, 1000\}$ on Adult 10k (QI_{demo}). The aggregate plateau of Section 3 conceals distinct per-feature dynamics.

Unlocking at $\varepsilon=1$: education-num rises from 0.0% to 33.6% at $\varepsilon=1$, then retreats to ~22–24% (MST’s pairwise model captures the education marginal more cleanly at lower noise but does not further improve). hours-per-week shows the same pattern (0.0 $\rightarrow$ 1.2%).

Noise artifact at $\varepsilon=1$: capital-gain and capital-loss briefly become non-zero at $\varepsilon=1$ (~1.0%) then return to 0% at $\varepsilon \geq 10$. At exactly $\varepsilon=1$, MST’s selected marginals include the capital features at low but nonzero signal; at higher ε , richer marginals displace them from the spanning tree.

Table 11: Reconstruction accuracy (R_{adv} , %) on CDC Diabetes. Mode baseline $\approx 42.5\%$. Bold = best per column among ML attacks. $\dagger$ = novel attack. (1k rows, QI_{demo} .)

Attack	RankSwap	Cell Supp.	Synthpop	MST ($\epsilon=0.1$)	MST ($\epsilon=1$)	MST ($\epsilon=10$)	MST ($\epsilon=100$)	MST ($\epsilon=1000$)	AIM ($\epsilon=1$)	AIM ($\epsilon=3$)	AIM ($\epsilon=10$)	TVAE	CTGAN	ARF	TabDDPM
Mode	42.5	42.5	42.5	41.7	42.5	42.5	42.5	42.5	42.5	42.5	42.5	42.5	42.5	42.5	42.5
KNN	83.5	59.4	46.7	41.5	42.2	44.1	44.2	44.2	43.1	43.1	45.1	45.1	42.6	46.7	75.7
Random Forest	75.6	58.6	46.9	41.5	42.5	43.9	44.1	44.2	42.6	43.5	45.0	45.2	42.4	45.0	75.1
MLP	61.5	57.3	48.4	41.9	43.1	44.9	45.4	45.2	43.2	43.9	45.7	46.2	42.8	46.6	71.0
CondDDPM $\dagger$	81.9	47.6	46.8	42.2	42.3	43.8	44.2	44.3	42.8	43.8	45.0	45.1	42.3	46.3	73.0
CoBP-RA $\dagger$	74.6	59.0	48.6	42.5	42.5	45.6	45.8	45.7	43.1	44.2	46.1	46.3	42.6	46.8	74.1

Table 12: Reconstruction accuracy (R_{adv} , %) on CDC Diabetes at 100k rows. Mode baseline = 42.2% throughout; bold = best per column among ML attacks. KNN, CondDDPM, and TabPFN were not run at this scale. $\dagger$ = novel attack. (QI_{demo} .)

Attack	RankSwap	Cell Supp.	Synthpop	MST ($\epsilon=0.1$)	MST ($\epsilon=1$)	MST ($\epsilon=10$)	MST ($\epsilon=100$)	MST ($\epsilon=1000$)	AIM ($\epsilon=1$)	TVAE	CTGAN	ARF	TabDDPM
Mode	42.2	42.2	42.2	42.2	42.2	42.2	42.2	42.2	42.2	42.2	42.2	42.2	42.2
Random Forest	62.9	87.1	47.3	43.6	43.8	43.3	43.2	43.3	45.6	45.1	45.2	45.8	47.7
MLP	45.9	46.9	45.6	43.7	43.2	43.2	43.2	43.2	46.0	45.0	45.2	45.0	45.9
Naive Bayes	49.2	49.5	48.4	45.1	46.5	47.2	47.3	48.1	48.9	48.1	48.2	48.8	49.6
CoBP-RA $\dagger$	62.9	85.7	47.8	44.4	44.0	44.1	44.0	44.0	46.0	45.7	46.1	46.5	48.1

Table 13: Per-feature R_{adv} (%) averaged over five attacks (NaiveBayes, KNN, RF, CoBP-RA, MLP) under MST. Bold = maximum per row. (Adult, 10k rows, QI_{demo} .)

Feature	$\epsilon=0.1$	$\epsilon=1$	$\epsilon=10$	$\epsilon=100$	$\epsilon=1000$
income	62.2	56.9	57.9	57.6	57.0
relationship	33.3	43.8	41.8	46.0	46.0
education-num	0.0	33.6	21.7	23.9	23.9
workclass	11.8	11.4	12.0	12.0	12.1
occupation	7.5	10.4	11.2	11.8	11.8
hours-per-week	0.0	1.2	1.1	1.2	1.2
capital-loss	0.0	1.2	0.0	0.0	0.0
capital-gain	0.0	1.0	0.0	0.0	0.0
fnlwt	0.0	0.0	0.0	0.0	0.0

Monotone riser: occupation (15 categories) is the only feature whose R_{adv} increases continuously ($7.5 \rightarrow 10.4 \rightarrow 11.2 \rightarrow 11.8\%$) throughout the full ϵ range. Its high cardinality means each incremental increase in privacy budget translates to marginally better conditional structure, unlike binary or low-cardinality features that saturate quickly.

Always zero: fnlwt (8,477 distinct values) is unrecoverable at all ϵ .

E.4 CDC Diabetes Reconstruction

Table 11 gives the full per-generator breakdown for CDC Diabetes (1k rows, QI_{demo}) summarized in Section 3.2. The dataset is the most exposed categorical setting we test under de-identification (KNN 83.5% on rank swapping), and TabDDPM nearly matches it (75.7%)

despite being fully synthetic. DP is correspondingly tight: across the entire MST sweep KNN moves only $41.5 \rightarrow 44.2\%$. CoBP-RA leads or ties most columns. Because the features are mostly binary the Mode baseline is high (42.5%), so it is the *gap above baseline*, not the raw value, that measures attack advantage.

E.5 Dataset Size: CDC Diabetes at Scale

Table 12 reports reconstruction on CDC Diabetes at 100,000 rows, the same attacks and generators as the 1k results in Table 11. The size effect is strongly mechanism-dependent (Section 3): de-identification grows *more* exposed with scale (Cell Suppression: RF $58.6\% \rightarrow 87.1\%$, as more QI groups clear the k -anonymity threshold and ship verbatim), while high-fidelity diffusion grows *less* exposed (TabDDPM: RF $75.1\% \rightarrow 47.7\%$, as more training data dilutes per-record memorization). MST remains pinned near its baseline at every budget.

E.6 QI Analysis: CDC Diabetes

Table 14 repeats on CDC Diabetes the QI analysis run for Adult (Section 3.3), and the headline holds: composition outweighs size. Enlarging the known set from 4 to 16 features lifts the random forest only +1.6 pp on the shared hidden-feature intersection, and every attack has saturated by the 10-feature QI_{demo} . Composition moves risk more: at a matched $|QI|=10$ the demographic QI beats the behavioral one by +4.0 pp for the random forest and +4.5 for KNN: the *opposite* sign to Adult, where behavioral knowledge dominated. On this mostly-binary health data the demographic attributes carry the predictive structure, whereas on Adult the employment and

Table 14: Reconstruction accuracy by quasi-identifier setting; the CDC counterpart of Table 6. Within each section, cells average over the *intersection* of hidden features across that section’s QI variants, so columns are comparable within a section (per-variant QI memberships in Table 24); bold = row maximum within a section. $\Delta_{\text{size}} = \text{QI}_{\text{large}} - \text{QI}_{\text{tiny}}$, $\Delta_{\text{comp}} = \text{QI}_{\text{beh.}} - \text{QI}_{\text{demo}}$, and $\Delta_{t \rightarrow d}^*$ is the gain from tiny to demographic QI, scored on the broader 12-feature two-way intersection. (CDC Diabetes, 1k rows, mean over all SDG methods.)

Attack	QI size sweep ($\nearrow$ adversary knowledge)					Composition contrast (matched QI)		
	QI _{tiny} (4 kn.)	QI _{demo} (10 kn.)	$\Delta_{t \rightarrow d}^*$	QI _{large} (16 kn.)	Δ_{size}	QI _{demo} (10 kn.)	QI _{beh.} (10 kn.)	Δ_{comp}
Mode	34.70	34.70	+0.00	34.70	+0.00	27.05	27.05	+0.00
Random	34.64	34.67	+0.01	34.65	+0.01	27.00	27.02	+0.02
KNN	40.12	41.98	+2.01	42.04	+1.91	34.18	29.69	−4.49
Naive Bayes	36.19	38.50	+2.36	38.70	+2.51	30.91	29.29	−1.62
Random Forest	39.50	40.98	+1.73	41.09	+1.59	33.28	29.24	−4.03
LightGBM	38.72	40.73	+2.19	40.89	+2.18	33.09	28.84	−4.25
MLP	38.78	40.33	+1.69	40.83	+2.06	32.63	29.27	−3.36

hours features did; in both, *which* features the adversary knows matters more than *how many*.

E.7 NIST Privacy CRC Scoreboard

Attack choice in the competition setting. Our winning QI₂ submissions used KNN and NaiveBayes rather than the tree ensembles that lead on Adult, a reminder that the best attack is scenario-dependent (Section 2). With a small training set ($n=10\,000$) and a high-dimensional ordinal QI (12 features), neighbor lookup effectively finds demographic twins, and NaiveBayes exploits strong independent QI–target marginals, whereas a random forest needs more data per leaf to learn the interactions that grow sparse as QI dimensionality rises. As a rule of thumb, small high-dimensional-QI settings favor KNN and NaiveBayes; larger datasets with moderate QIs favor RF and MLP.

E.8 Quasi-Identifier Definitions

Tables 23–27 enumerate features and quasi-identifier memberships for each of the five datasets evaluated in this paper. For Adult and CDC Diabetes (Tables 23–24), features are shown as columns with a $\checkmark$ marking membership in each QI variant; blank = hidden. NIST Arizona (Table 25) uses the same format for the two competition QI variants, QI₁ and QI₂. California Housing (Table 26) has two QI variants. For NIST SBO (Table 27), the 130 features are too numerous for a column-per-feature layout; instead, features are grouped into 16 semantic categories and QI membership is described per category. QI_{demo} for Adult/CDC/Arizona corresponds to QI1 in the source code.

E.9 Continuous-Domain Memorization

The memorization test of Section 4.1 carries over to continuous data by replacing rarity-weighted accuracy with normalized RMSE (lower is better) and a regression attack for the classifier. We run

three regressors (a random forest, Bayesian ridge regression, and the continuous variant of CoBP-RA, Appendix C.3) on California Housing and report, for each, the training-target error tr and the train-minus-holdout gap Δ (Table 16). The same three-regime structure as the categorical datasets reappears, with the memorization sign reversed by the change of metric: resynthesizing generators that memorize (TabDDPM, Synthpop) show negative Δ , while the DP generators sit at $\Delta \approx 0$.

One pattern is worth isolating, because it shows that *reconstruction accuracy and memorization sensitivity are not the same quantity*. On TabDDPM and Synthpop, the continuous CoBP-RA has by far the *worst* training error ($\text{tr} = 0.158$ and 0.191 , versus 0.067 and 0.107 for the random forest) yet by far the *largest* memorization gap ($\Delta = -0.112$ and -0.115 , versus -0.034 and -0.033). A smooth global regressor like Bayesian ridge, conversely, reconstructs reasonably but registers essentially no gap ($\Delta \approx -0.001$). The explanation is what each attack keys on: CoBP-RA’s per-record neighbourhood couplings (local PMI from a target’s nearest synthetic rows) latch onto exactly the idiosyncratic joint structure a memorizing generator copies from individual training records, so they fire much harder on members than on holdouts even though the resulting point estimate is noisier overall. Bayesian ridge, fitting one global linear map, cannot express that record-specific structure and so neither reconstructs it nor detects it. The practical lesson for auditing: the best memorization *detector* need not be the best *reconstructor*; an attack that exploits local joint structure can expose memorization that a lower-error but globally-smooth attack completely misses.

E.10 Disparate Impact: Per-Attack and Per-Feature

These tables support the claim in Section 4.3 that disparate exposure is set by the *generator*, not by the attack, and that it concentrates on features whose values are individually identifying. Both use the

Team	Mean over all SDG		Per-SDG score (QI_1)								
	QI_1 (avg.)	QI_2 (avg.)	Cell Supp.	Rank Swap	ARF	TVAE	Synth- pop	MST ($\epsilon=1$)	MST ($\epsilon=10$)	AIM ($\epsilon=1$)	AIM ($\epsilon=10$)
(ours)	654	598	415	1078	346	343	363	349	338	351	349
(ours)	569	506	414	438	317	342	365	314	307	290	314
(ours)	573	499	430	426	324	343	365	325	306	293	325
?	381	373	248	245	243	246	245	245	242	244	245
?	370	359	246	245	241	243	246	228	222	220	228

Table 15: Official NIST Privacy CRC scoreboard. Rows = competing teams; our three submissions appear in the first three rows. R_{adv} (Eq. 1) is *summed* across all hidden features per the official scoring protocol. The Mean over all SDG columns give each submission’s average over all nine SDG methods, under known- QI sets QI_1 (7 features, 18 hidden) and QI_2 (12 features, 10 hidden); the nine right-hand columns are per-SDG scores under QI_1 . Feature-level QI membership is in Table 25. (NIST Arizona 1940 census, 25 features, $n=10,000$.)

Table 16: Memorization test on California Housing. NRMSE: lower = better reconstruction; Δ = $NRMSE_{tr} - NRMSE_{NT}$, so *negative* Δ signals memorization (bold: $\Delta \leq -0.015$). $\ddagger$ At MST $\epsilon=0.1$ the implied holdout NRMSE ($tr - \Delta$) exceeds the mean-imputation baseline (≈ 0.24) for all three attacks, indicating attack failure on very noisy data rather than memorization. $\dagger$ = novel attack. (QI_{large} , 1k rows.)

SDG method	RandomForest		BayesianRidge		CoBP-RA †	
	tr	Δ	tr	Δ	tr	Δ
TabDDPM	0.067	-0.034	0.119	-0.001	0.158	-0.112
Synthpop	0.107	-0.033	0.121	-0.001	0.191	-0.115
RankSwap	0.092	-0.017	0.120	-0.002	0.245	-0.007
ARF	0.120	-0.005	0.132	-0.001	0.204	+0.001
TVAE	0.136	-0.002	0.147	+0.001	0.194	-0.001
CTGAN	0.172	+0.001	0.165	+0.001	0.245	+0.022
MST $\epsilon=0.1^\ddagger$	0.264	-0.026	0.232	-0.008	0.317	-0.123‡
MST $\epsilon=1$	0.183	0.000	0.184	+0.001	0.182	0.000
MST $\epsilon=10$	0.154	-0.002	0.156	-0.004	0.169	0.000
MST $\epsilon=100$	0.154	0.000	0.160	-0.003	0.166	+0.002
MST $\epsilon=1000$	0.158	-0.007	0.153	-0.007	0.252	-0.039

same setting as Table 9 (Adult, 10k, QI_{demo}); outliers are the top-10% by QI -space Isolation Forest.

E.11 Data Qualities

Tables 19–22 extend the main-text fidelity/utility overview (Table 2, Adult) to the remaining four datasets, using the metric definitions of Table 2 together with one additional column, **Prop. AUC** (propensity AUC: the AUC of a classifier trained to distinguish real from synthetic records; $\downarrow$, 0.5 = indistinguishable), and the same favorable-direction arrows; they let the reconstruction results of Section 3 be read against the quality of the release each attack faced. The recurring pattern is that the de-identification methods (RankSwap, CellSuppression) lead most fidelity metrics because they retain real records (which is also what makes them the most exposed), while the DP generators trade fidelity for their ϵ guarantee. The four datasets sharpen this differently. On **NIST Arizona** (Table 19) TabDDPM is the strongest *synthetic* generator (TSTR ratio 1.02)

Table 17: Per-attack outlier penalty: each cell is the *ratio* of mean row-level R_{adv} on the outlier 10% to that on the typical 90% (>1 = atypical individuals reconstructed more accurately), with the absolute mean row R_{adv} (%) in parentheses: the Table 9 penalty broken out by attack $\times$ generator. The ratio is meaningful only where the attack reconstructs: cells with mean row $R_{adv} > 1.5\%$ are bold, and ratios on near-zero- R_{adv} cells (noise over noise) should be ignored. $\dagger$ = novel attack. (Adult, 10k rows, QI_{demo} .)

Attack	CellSupp.	TabDDPM	Synthpop	MST ($\epsilon=1$)
Mode	1.05 (0.2)	0.73 (0.1)	0.67 (0.1)	0.78 (0.1)
KNN	2.18 (34.5)	2.88 (1.6)	0.79 (1.5)	0.76 (0.1)
NaiveBayes	2.18 (32.1)	3.61 (1.0)	1.31 (1.7)	0.83 (0.1)
RandomForest	2.17 (34.7)	2.56 (1.7)	0.92 (1.9)	0.94 (0.1)
CoBP-RA †	2.17 (34.7)	2.62 (1.7)	1.15 (1.8)	0.92 (0.1)
MLP	2.97 (0.6)	2.79 (0.5)	1.14 (1.9)	0.96 (0.1)
LightGBM	2.96 (0.9)	1.51 (0.2)	1.65 (0.4)	0.91 (0.1)

and MST fidelity climbs steadily with ϵ . On **CDC Diabetes** (Table 20), whose features are mostly binary, fidelity is high across the board and three generators saturate the TSTR-ratio cap. On **NIST SBO** (Table 21) the 130 noisy financial columns invert the usual order: TabDDPM collapses (Col. pairs 0.35) while RankSwap and the higher- ϵ MST models fare best. And on **California Housing** (Table 22), fully continuous, the DP marginal models lose most utility (TSTR $R^2 \rightarrow 0$ at small ϵ) whereas TabDDPM and RankSwap retain it.

Table 18: Per-feature outlier gap for the two leaking generators: each cell is mean row-level R_{adv} on the outlier 10% *minus* that on the typical 90%, averaged over the attacks that reconstruct, so a positive value means the feature is reconstructed *better* for atypical individuals, localizing the Table 17 penalty to specific features. The gap is positive on individually identifying features (occupation, hours-per-week) and negative on features dominated by one common value (education-num, relationship). fnlwgt is the sharpest mechanism contrast: record retention exposes it for outliers (CellSupp. +0.36) but synthesis cannot reproduce a near-unique value even for them (TabDDPM +0.02). (Adult, 10k rows, QI_{demo} .)

Feature	Type	CellSupp.	TabDDPM
fnlwgt	cont. (near-unique)	+0.36	+0.02
occupation	categorical	+0.22	+0.21
hours-per-week	numeric	+0.22	+0.18
income	binary	+0.11	+0.11
workclass	categorical	+0.07	+0.06
capital-gain	cont. (sparse)	+0.07	+0.04
capital-loss	cont. (sparse)	+0.04	+0.05
education-num	ordinal	-0.05	-0.04
relationship	categorical	-0.08	-0.07

Table 19: Synthetic-data quality profile for NIST Arizona; metric definitions as in Table 2, best per column excluding *Train–Train* in bold. (1940 census, 25 features, 10k rows.)

SDG method	TSTR ratio $\uparrow$	Mean JSD $\downarrow$	Pairwise TVD $\downarrow$	Wass. OHE $\downarrow$	Corr. diff $\downarrow$	Col. pairs $\uparrow$	Prop. AUC $\downarrow$
MST ($\varepsilon=0.1$)	0.711	0.101	0.140	5.787	0.157	0.773	0.858
MST ($\varepsilon=1$)	0.778	0.036	0.048	3.117	0.061	0.927	0.932
MST ($\varepsilon=10$)	0.719	0.010	0.026	2.553	0.046	0.968	0.960
MST ($\varepsilon=100$)	0.770	0.005	0.023	2.497	0.042	0.975	0.959
MST ($\varepsilon=1000$)	0.774	0.004	0.023	2.495	0.042	0.975	0.965
AIM ($\varepsilon=1$)	0.796	0.054	0.055	4.573	0.064	0.890	0.588
RankSwap	1.041	0.000	0.000	0.023	0.035	0.966	0.451
Cell Supp.	1.045	0.012	0.013	0.365	0.007	0.977	0.517
Synthpop	0.986	0.010	0.033	0.297	0.009	0.919	0.499
TVAE	0.897	0.094	0.129	3.711	0.039	0.827	0.599
CTGAN	0.712	0.090	0.133	4.415	0.053	0.829	0.655
ARF	0.961	0.064	0.128	3.202	0.036	0.798	0.561
TabDDPM	1.019	0.020	0.024	0.627	0.022	0.963	0.506
<i>Train–Train</i>	–	0.013	0.019	–	0.009	0.977	–

Table 20: Synthetic-data quality profile for CDC Diabetes; metric definitions as in Table 2, best per column excluding *Train–Train* in bold. (22 features, mostly binary, 1k rows.)

SDG method	TSTR ratio $\uparrow$	Mean JSD $\downarrow$	Pairwise TVD $\downarrow$	Wass. OHE $\downarrow$	Corr. diff $\downarrow$	Col. pairs $\uparrow$	Prop. AUC $\downarrow$
MST ($\varepsilon=0.1$)	0.762	0.210	0.386	13.083	0.193	0.539	0.983
MST ($\varepsilon=1$)	0.902	0.057	0.105	4.630	0.126	0.793	0.823
MST ($\varepsilon=10$)	0.916	0.007	0.033	4.022	0.133	0.939	0.575
MST ($\varepsilon=100$)	0.910	0.002	0.025	3.976	0.087	0.955	0.539
MST ($\varepsilon=1000$)	0.886	0.001	0.026	4.043	0.069	0.955	0.548
AIM ($\varepsilon=1$)	0.898	0.063	0.099	4.544	0.193	0.770	0.785
RankSwap	1.100	0.000	0.000	0.010	0.017	0.989	0.485
Cell Supp.	1.100	0.083	0.141	3.603	0.087	0.834	0.749
Synthpop	1.053	0.014	0.031	0.700	0.034	0.958	0.473
TVAE	0.994	0.137	0.198	5.637	0.073	0.720	0.848
CTGAN	0.892	0.069	0.141	4.304	0.197	0.779	0.778
ARF	0.987	0.023	0.047	1.590	0.035	0.891	0.524
TabDDPM	1.100	0.019	0.034	0.941	0.031	0.953	0.501
<i>Train–Train</i>	–	0.018	0.037	–	0.039	0.948	–

Table 21: Synthetic-data quality profile for NIST SBO; metric definitions as in Table 2, best per column excluding *Train-Train* in bold. Pairwise TVD is omitted (not computed for >30 categorical columns); the large Wasserstein values reflect SBO’s noisy continuous financial variables dominating the OHE-space distance. (130 features, financial, 1k rows.)

SDG method	TSTR ratio $\uparrow$	Mean JSD $\downarrow$	Wass. OHE $\downarrow$	Corr. diff $\downarrow$	Col. pairs $\uparrow$	Prop. AUC $\downarrow$
MST ($\epsilon=0.1$)	0.979	0.407	129.4	0.263	0.277	0.999
MST ($\epsilon=1$)	1.013	0.103	125.5	0.275	0.864	0.955
MST ($\epsilon=10$)	0.992	0.029	122.1	0.177	0.952	0.799
MST ($\epsilon=100$)	0.990	0.006	121.9	0.175	0.971	0.917
MST ($\epsilon=1000$)	0.997	0.004	122.0	0.141	0.974	0.958
RankSwap	1.100	0.000	0.004	0.073	1.000	0.480
Cell Supp.	1.004	0.059	9.966	0.087	0.959	0.612
Synthpop	1.011	0.019	3.604	0.051	0.975	0.486
TVAE	0.989	0.160	30.919	0.144	0.852	0.946
CTGAN	0.994	0.118	27.380	0.270	0.837	0.799
ARF	1.003	0.027	7.286	0.073	0.966	0.603
TabDDPM	1.027	0.423	122.0	0.486	0.353	0.997
<i>Train-Train</i>	–	0.021	–	0.088	0.978	–

Table 22: Synthetic-data quality profile for California Housing; metric definitions as in Table 2, best per column excluding *Train-Train* in bold. TSTR ratio uses regression R^2 (not F1-macro); the categorical-marginal metrics (mean JSD, pairwise TVD) are undefined for fully continuous data and omitted, and Col. pairs is computed over discretized columns. (9 continuous features, 1k rows.)

SDG method	TSTR ratio $\uparrow$	Wass. OHE $\downarrow$	Corr. diff $\downarrow$	Col. pairs $\uparrow$	Prop. AUC $\downarrow$
MST ($\epsilon=0.1$)	0.000	10.275	0.174	0.610	0.977
MST ($\epsilon=1$)	0.000	15.043	–	0.692	0.947
MST ($\epsilon=10$)	0.484	13.464	0.124	0.760	0.870
MST ($\epsilon=100$)	0.610	13.989	0.121	0.795	0.815
MST ($\epsilon=1000$)	0.586	14.663	0.101	0.903	0.933
AIM ($\epsilon=1$)	0.000	14.956	–	0.595	0.901
RankSwap	1.100	0.002	0.047	0.914	0.488
Synthpop	1.046	1.038	0.036	0.955	0.495
TVAE	0.708	5.828	0.121	0.706	0.913
CTGAN	0.076	7.757	0.167	0.617	0.945
ARF	0.854	2.434	0.050	0.888	0.547
TabDDPM	1.100	0.990	0.043	0.953	0.475
<i>Train-Train</i>	–	–	0.043	0.965	–

Table 23: QI membership for the *Adult* dataset (15 features, 4 QI variants). ✓ = known to adversary (quasi-identifier); blank = hidden (must be reconstructed). Features are grouped by the tier at which they first enter a QI.

QI ($ \text{QI} $)	tiny core			+ QI _{demo}			+ QI _{large}				beh. only		hidden		
	age	sex	race	education	marital-status	native-country	occupation	workclass	relationship	hours-per-week	education-num	fnlwgt	capital-gain	capital-loss	income
QI _{tiny} (3)	✓	✓	✓												
QI _{demo} (6)	✓	✓	✓	✓	✓	✓									
QI _{large} (10)	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓					
QI _{beh.} (6)							✓	✓	✓	✓	✓	✓			

Table 24: QI membership for the *CDC Diabetes* dataset (22 features, 4 QI variants). ✓ = known to adversary; blank = hidden. Features *Smoker* and *PhysActivity* (grouped under “+ demo”) also appear in QI_{beh.}; similarly, the six “+ large” features all overlap with QI_{beh.}. Abbreviated names: *HvyAlcohol* = *HvyAlcoholConsump*; *HeartDisease* = *HeartDiseaseorAttack*.

QI ($ QI $)	tiny core				+ QI _{demo}					+ QI _{large}					beh. only		hidden					
	Sex	Age	BMI	GenHlth	Education	Income	HighBP	HighChol	Smoker	PhysActivity	Fruits	Veggies	CholCheck	DiffWalk	AnyHealthcare	NoDocbcCost	HvyAlcohol	HeartDisease	Stroke	Diabetes_binary	MentHlth	PhysHlth
QI _{tiny} (4)	✓	✓	✓	✓																		
QI _{demo} (10)	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓												
QI _{large} (16)	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓						
QI _{beh.} (10)									✓	✓	✓	✓	✓	✓	✓	✓	✓	✓				

Table 25: QI membership for the *NIST Arizona* dataset, 25-feature competition subset (25 IPUMS features, 2 QI variants). ✓ = known to adversary; blank = hidden. QI₁ captures cultural/identity attributes; QI₂ captures geographic/mobility attributes. Both share RACE and SEX. These are the QI sets used in the NIST Privacy CRC competition (25-feature version). QI_{medium} is used in the RA-as-MIA experiment described in Section 4.2.

	both QIs		QI ₁ only					QI ₂ only					hidden												
	RACE	SEX	AGEMARR	GQTYPE	IND	MTONGUE	VETSTAT	BPL	FARM	HISPAN	LABFORCE	MIGRATE5	AGE	CITIZEN	DURUNEMP	EDUC	EMPSTAT	FAMSIZE	GQ	INCWAGE	MARST	NATIVITY	OWNERSHP	URBAN	WKSWORK1
QI ₁ (7)	✓	✓	✓	✓	✓	✓	✓																		
QI ₂ (7)	✓	✓						✓	✓	✓	✓	✓													
QI _{medium} (12)	✓	✓	✓	✓	✓	✓	✓	✓		✓	✓		✓				✓								

Table 26: QI membership for the *California Housing* dataset (9 continuous features, 2 QI variants). ✓ = known to adversary; blank = hidden (must be reconstructed). QI_1 : geographic and locational features only. QI_{large} : adds structural housing features; hides only the three economic outcomes.

QI ($ QI $)	QI_1				+ QI_{large}		hidden	
	Latitude	Longitude	HouseAge	Population	AveBedrms	AveRooms	MedInc	AveOccup
QI_1 (4)	✓	✓	✓	✓				
QI_{large} (6)	✓	✓	✓	✓	✓	✓		

Table 27: QI membership for the *NIST SBO* dataset (130 features in 16 semantic categories). Under QI_1 (the single QI variant evaluated for SBO; QI_1 in the source code) the adversary knows 7 firm-identifier and owner-demographic features, listed in the first two rows; the remaining 123 are hidden. Features are grouped by category for readability, and a dash (–) marks a category with no features known under QI_1 .

Feature category	N	Known (QI_1 , 7)	Hidden (123)
Business identifiers	6	FIPST, SECTOR, N07_EMPLOYER, RG	TABWGT, PCT1
Owner demographics	4	SEX1, RACE1, ETH1	VET1
Business acquisition	6	–	FOUNDED1, PURCHASED1, INHERITED1, RECEIVED1, ACQUIRENR1, ACQYR1
Owner role & background	12	–	PROVIDE1, MANAGE1, FINANCIAL1, FNCTNABV1, FNCTNR1, HOURS1, PRMNC1, SELFEMP1, EDUC1, AGE1, BORNUS1, DISVET1
Basic business	4	–	ESTABLISHED, HOMEBASED, FRANCHISE, FRANCHISER50
Customer / market type	8	–	FEDERAL, STATELOCAL, OTHERBUS, INDIVIDUALS, CUSTNR, EXPORTS, OPSOUTSIDE, OUTSOURCE
Languages spoken	18	–	ENGLISH, ARABIC, CHINESE, FRENCH, GERMAN, GREEK, HINDI, ITALIAN, JAPANESE, KOREAN, POLISH, PORTUGUESE, RUSSIAN, SPANISH, TAGALOG, VIETNAMESE, LANGOTHER, LANGNR
Workforce composition	7	–	FULLTIME, PARTTIME, DAYLABOR, TEMPSTAFF, LEASED, CONTRACTORS, EMPNR
Digital presence	4	–	WEBSITE, ECOMMERCE, ECOMMPCT, ONLINEPURCH
Operating schedule / activity	6	–	LT40HOURS, LT12MONTHS, SEASONAL, OCCASIONALLY, ACTIVITYNABV, ACTIVITYNR
Ownership structure	3	–	HUSBWIFE, FAMILYBUS, NUMOWNERS
Financial outcomes	3	–	EMPLOYMENT_NOISY, PAYROLL_NOISY, RECEIPTS_NOISY
Startup capital access (SC*)	15	–	SCSAVINGS, SCASSETS, SCEQUITY, SCCREDIT, SCGOVTLOAN, SCGOVTGUAR, SCBANKLOAN, SCFAMLOAN, SCVENTURE, SCGRANT, SCOTHER, SCDONTKNOW, SCNONENEDED, SCNOTREPORTED, SCAMOUNT
Expansion capital access (EC*)	16	–	ECSAVINGS, ECASSETS, ECEQUITY, ECCREDIT, ECGOVTLOAN, ECGOVTGUAR, ECBANKLOAN, ECFAMLOAN, ECVENTURE, EC PROFITS, ECGRANT, ECOTHER, ECDONTKNOW, ECNOACCESS, ECNOEXPAND, ECNOTREPORTED
Employee benefits	6	–	HEALTHINS, RETIREMENT, PROFITSHARE, HOLIDAYS, BENENABV, BENENR
Business cessation	12	–	OPERATING, RETIRED, DECEASED, ONETIME, LOWSALES, NOBUSCRED, NOPERSCRED, STARTANOTHER, SOLDBUS, CEASEOTHER, CEASENR, CEASENA
<i>Total</i>	<i>130</i>	<i>7</i>	<i>123</i>